2016 No. 017

Independent Evaluation of the Validity and Reliability of STAAR Grades 3-8 Assessment Scores: Part 2

Final Report

Prepared for:	Texas Education Agency Student Assessment Division William B. Travis Building 1701 N. Congress Avenue Austin, Texas, 78701	Prepared under:	Contract # 3436
Prepared by:	Human Resources Research Organization (HumRRO)	Date:	April 28, 2016

Independent Evaluation of the Validity and Reliability of STAAR Grades 3-8 Assessment Scores: Part 2

Table of Contents

List of Tables

Independent Evaluation of the Validity and Reliability of STAAR Grades 3-8 Assessment Scores: Part 2

Executive Summary

The Texas Education Agency (TEA) contracted with the Human Resources Research Organization (HumRRO) to provide an independent evaluation of the validity and reliability of the State of Texas Assessments of Academic Readiness (STAAR) scores, including grades 3-8 reading and mathematics, grades 4 and 7 writing, grades 5 and 8 science, and grade 8 social studies. The independent evaluation is intended to support HB 743, which states that before an assessment may be administered, “the assessment instrument must, on the basis of empirical evidence, be determined to be valid and reliable by an entity that is independent of the agency and of any other entity that developed the assessment instrument.” Our independent evaluation consists of three tasks that are intended to provide empirical evidence for both the validity of the STAAR scores (Task 1) and for the projected reliability of the assessment (Task 2). Validity and reliability are built into an assessment by ensuring the quality of all of the processes employed to produce student test scores. Under Task 3, we reviewed the procedures used to build and score the assessment. The review focuses on whether the procedures support the creation of valid and reliable assessment scores.

HumRRO’s independent evaluation finds support for the validity and reliability of the 2016 STAAR scores. Specifically:

- Under Task 1, we identified evidence of the content validity of the assessments. The content review consisted of rating the alignment of each item to the Texas Essential Knowledge and Skills (TEKS) expectation the item was intended to measure. Overall, the content of the 2016 forms aligned with blueprints and the vast majority of items were aligned with the TEKS expectations for grades 3 through 8 mathematics and reading, grades 5 and 8 science, grade 8 social studies, and grades 4 and 7 writing.
- Our work associated with Task 2 provided empirical evidence of the projected reliability and standard error of measurement for the 2016 forms. The projected reliability and conditional standard error of measurement (CSEM) estimates were all acceptable. Assuming the 2016 students’ scores will have a similar distribution as the 2015 scores and assuming similar item functioning, the reliability and CSEM estimates based on 2016 student data should be similarly acceptable.
- Finally, under Task 3, we reviewed the documentation of the test construction and scoring processes. Based on HumRRO’s 20 years of experience in student achievement testing and 30 years of experience in high-stakes test construction, the processes used to construct the 2016 tests and the proposed methods for scoring the 2016 test are consistent with industry standards and support the development of tests that measure the knowledge and skills outlined in the content standards and test blueprint. The processes allow for the development of tests that yield valid and reliable assessment scores.

Independent Evaluation of the Validity and Reliability of STAAR Grades 3-8 Assessment Scores: Part 2

The Texas Education Agency (TEA) contracted with the Human Resources Research Organization (HumRRO) to provide an independent evaluation of the validity and reliability of the State of Texas Assessments of Academic Readiness (STAAR) scores, including grades 3-8 reading and mathematics, grades 4 and 7 writing, grades 5 and 8 science, and grade 8 social studies. The independent evaluation is intended to support HB 743, which states that before an assessment may be administered, “the assessment instrument must, on the basis of empirical evidence, be determined to be valid and reliable by an entity that is independent of the agency and of any other entity that developed the assessment instrument.” Our independent evaluation consists of three tasks that are intended to provide empirical evidence for both the validity of the STAAR scores (Task 1) and for the projected reliability of the assessment (Task 2). Validity and reliability are built into an assessment by ensuring the quality of all of the processes employed to produce student test scores. Under Task 3, we reviewed the procedures used to build and score the assessment. The review focuses on whether the procedures support the creation of valid and reliable assessment scores.

This report includes results of the content review of the 2016 STAAR forms, projected reliability and standard error of measurement estimates for the 2016 STAAR forms, and a review of the processes used to create, administer, and score STAAR. Part 2 of the report expands upon results presented in Part 1 and includes results for mathematics and reading grades 3 through 8, science grades 5 and 8, social studies grade 8, and writing grades 4 and 7.

Overview of Validity and Reliability

Validity

Over the last several decades, testing experts from psychology and education[1] have joined forces to create standards for evaluating the validity and reliability of assessment scores, including those stemming from student achievement tests such as the STAAR. The latest version of the standards was published in 2014. Perhaps more applicable to Texas is the guidance given to states by the US Department of Education, which outlines requirements for the peer review of their student assessment programs.[2] The peer review document is, in essence, a distillation of several relevant parts of the AERA/APA/NCME guidelines. The purpose of this report is not to address all of the requirements necessary for peer review. That is beyond the scope of HumRRO’s contract. Rather, we are addressing the Texas Legislature’s requirement to provide a summary judgement about the assessment prior to the spring administrations. To that end, and to keep the following narrative accessible, we begin by highlighting a few relevant points related to validity and reliability.

“Validity” among testing experts concerns the legitimacy or acceptability of the interpretation and use of ascribed test scores. Validity is not viewed as a general property of a test because scores from a particular test may have more than one use. The major implication of this statement is that a given test score could be “valid” for one use but not for another. Evidence may exist to support one interpretation of the score but not another. This leads to the notion that

[1] A collaboration between the American Educational Research Association (AERA), American Psychological Association (APA), and the National Council on Measurement in Education (NCME).

[2] www2.ed.gov/admins/lead/account/peerreview/assesspeerrevst102615.doc

test score use(s) must be clearly specified before any statement can be made about validity. Thus, HumRRO began its validity review by simply listing the uses ascribed to STAAR in technical documents available from the TEA.

HumRRO reviewed on-line documents, including *Interpreting Assessment Reports: State of Texas Assessments of Academic Readiness (STAAR®) Grades 3-8*[3] and Chapter 4 of the *2014-2015 Technical Digest,*[4] to identify uses for STAAR scores for individual students. Three validity themes were identified:

1. STAAR grade/subject[5] scores are intended to be representative of what a student knows and can do in relation to that specific grade and subject. This type of validity evidence involves demonstrating that each grade/subject test bears a strong association with on-grade curriculum requirements, as defined by TEA standards and blueprints, for that grade and subject.
2. STAAR grade/subject scores, when compared to scores for a prior grade, are intended to be an indication of how much a student has learned since the prior grade.
3. STAAR grade/subject scores are intended to be an indication of what students are likely to achieve in the future.

For the purposes of our review, we focused on the first validity theme listed above, which is specific to the interpretation of on-grade STAAR scores for individual students. Validity evidence associated with interpreting growth (theme 2) or for projecting anticipated progress (theme 3) is outside the scope of this review.

Under Task 1, HumRRO conducted a content review to examine the content validity of the 2016 grades 3-8 STAAR test forms. Specifically, this review sought to determine how well the 2016 STAAR test forms align with the on-grade curriculum, as defined by the Texas content standards and assessment blueprints. Under Task 3, we reviewed test-building procedures to assess the extent to which the processes support intended test score interpretations.

Reliability

"Reliability" concerns the repeatability of test scores, and like validity, it is not a one-size-fits-all concept. There are different kinds of reliability – and the most relevant kind of reliability for a test score depends on how that score is to be used. Internal consistency reliability is an important consideration and the kind of reliability that is typically analyzed for large-scale educational assessment scores. This kind of test score reliability estimates how well a particular collection of test items relate to each other within the same theoretical domain. To the extent that a set of items is interrelated, or similar to each other, we can infer that other collections of related items would be likewise similar. That is, can we expect the same test score if the test contained a different set of items that were constructed in the same way as the given items?

[3] http://tea.texas.gov/student.assessment/interpguide/

[4] http://tea.texas.gov/Student_Testing_and_Accountability/Testing/Student_Assessment_Overview/Technical_Digest_2014-2015/

[5] We use the term "grade/subject" to mean any of the tested subjects for any of the tested grades (e.g., grade 4 mathematics or grade 5 science).

Another concept related to reliability is standard error of measurement (SEM). The technical term standard error of measurement refers to the notion that a test score cannot be perfect, and that every test score contains some degree of uncertainty. SEMs are computed for the entire range of test scores whereas *conditional* standard errors of measurement (CSEM) vary depending on each possible score. For example, if test items are all difficult, those items will be good for reducing uncertainty in reported scores for high achieving students, but will not be able to estimate achievement very well for average and below average students (who will all tend to have similar low scores). Small CSEM estimates indicate that there is less uncertainty in student scores. Estimates can be made at each score point and across the distribution of scores.

Internal consistency reliability and SEM estimates cannot be computed for a test until student response data are available. However, we can make projections about the reliability and SEM using the item response theory (IRT) parameter estimates that were used to construct test forms and projections of the distribution of student scores. To the extent that the items function similarly in 2016 to previous administrations and the 2016 STAAR student score distribution is similar to the 2015 STAAR score distribution, the projected reliability and SEM estimates should be very similar to those computed after the test administrations. A summary of these analyses is presented under the Task 2 heading.

Task 1: Content Review

HumRRO conducted a content review of the STAAR program to investigate the content validity of scores for grades 3-8 assessments. Specifically, this review sought to determine how well the items on the 2016 STAAR forms represented the content domain, defined by the content standard documents and test blueprints. This review included the 2016 assessments forms, standards documentation, and blueprints for mathematics and reading grades 3 through 8, science grades 5 and 8, social studies grade 8, and writing grades 4 and 7. The intent of this review was not to conduct a full alignment study. To comply with the peer review requirements, another contractor conducted a full alignment study of the STAAR program.

Background Information

HumRRO used three main pieces of documentation for each grade and content area to conduct the content review: (a) eligible Texas Essential Knowledge and Skills for each assessment[6], (b) assessment blueprints[7], and (c) 2016 assessment forms.

The Texas STAAR program measures the Texas Essential Knowledge and Skills (TEKS) for each grade and content area. The knowledge and skills are categorized by three or four reporting categories, depending on the content area. These reporting categories are general and consistent across grade levels for a given subject. There are one or more grade-specific knowledge and skills statements under each reporting category. Each knowledge and skill statement includes one or more expectations. The expectations are the most detailed level and describe the specific skills or knowledge students are expected to have mastered. Test items are written at the expectation level. Each expectation is defined as either a readiness or supporting standard. Texas defines readiness standards as those most pertinent for success in the current grade, and important for future course preparation. Supporting standards are those introduced in a previous grade or emphasized more fully in a later grade, but still important for the current grade.

The assessment blueprints provide a layout for each test form. For each grade/subject, the blueprints describe the number of items that should be included for each reporting category, standard type (readiness or supporting), and item type, when applicable. The blueprints also link back to the content standards documents by indicating the number of standards written to each reporting category and for the overall assessment.

Each assessment form includes between 19 and 56 items, depending on the grade and content area. The forms mostly include multiple choice items, with a few gridded items for mathematics and science, and one composition item for writing. The reading and social studies assessments include only multiple-choice items. Each item was written to a specific TEKS expectation. The forms follow the blueprint for distribution of items across reporting category, standards type, and item type.

[6] For Math, http://ritter.tea.state.tx.us/rules/tac/chapter111/index.html;
For Reading, http://ritter.tea.state.tx.us/rules/tac/chapter110/index.html

[7] http://tea.texas.gov/student.assessment/staar/#G_Assessments

Method

HumRRO reviewed two key pieces of evidence to examine how well the 2016 STAAR forms aligned to the content intended by the TEA. First, HumRRO determined how well the item distribution matched that specified in the assessment blueprints. Second, an alignment review was conducted to determine the extent to which each item was aligned to the intended TEKS student expectation.

To determine how well the test forms represented the test blueprint, the number of items falling within each reporting category, standard type, and item type (as indicated by the TEKS code) were calculated. These numbers were compared to the number indicated by the assessment blueprints.

To conduct the alignment review all items from each test form were rated by four HumRRO reviewers - with the exception of mathematics grades 3, 4, 6, and 7, where three reviewers rated each item. Each group of reviewers included those who had previous experience conducting alignment or item reviews and/or those with relevant content knowledge. All reviewers attended web-based training prior to conducting ratings. The training provided an overview of the STAAR program, background information about the TEA standards, and instructions for completing the review. Reviewers reviewed each item and the standard assigned to it. They assigned each item a rating of “fully aligned,” “partially aligned,” or “not aligned” to the intended standard. Ratings were made at the expectation level.

- A rating of “fully aligned” required that the item fully fit within the expectation.
- A rating of “partially aligned” was assigned if some of the item content fell within the expectation, but some of the content fell outside.
- A rating of “not aligned” was assigned if the item content fell outside the content included in the expectation.

A partial alignment rating should not be interpreted as misalignment; rather, a partially aligned item is one that includes some content of the intended TEKS expectation, but with some additional skills/knowledge required. For reading, the TEKS expectations specified genres, and in some cases, reviewers selected a partial alignment rating when they felt the passage for the item fit better in a different genre. While all reviewers were trained to assign ratings using the same methodology, a certain level of subjective judgement is required. We include information about the number of reviewers who assigned “partially aligned” or “not aligned” ratings for each grade at each reporting category to provide perspective. Item level information, including reviewer justification, for items rated partially or not aligned is provided in an addendum.

In addition to these ratings, if a reviewer provided a rating of “partially aligned” or “not aligned” he or she was asked to provide information about what content of the item was not covered by the aligned expectation and, if appropriate, to provide an alternate expectation to which the item better aligned.

During training reviewers were given the opportunity to practice assigning ratings for a selection of items. At this time, the HumRRO content review task lead ensured all reviewers properly understood how to use the rating forms and standards documentation, and how to apply ratings. Once completed, ratings were reviewed to ensure the reviewers were interpreting the process consistently and appropriately. If there were specific questions about a rating, the content review task lead discussed the issue with the reviewer to determine the most appropriate course of action. If reviewers' interpretations were inconsistent with the methodology, ratings were revised.

To obtain the average percentage of items at each alignment level (full, partial, or not) the following steps were taken:

1. Determine the percentage of items fully, partially, or not aligned to the intended TEKS expectation for each reviewer; and
2. Average the percentages across reviewers.

Therefore, the percentages reported take into account all individual ratings and are averages of averages. As an example, to get the average percentage of items "partially aligned" for a reporting category, the following calculation is used:

$$Average\ \% = \frac{\sum_{k=1}^{K}\left(\frac{\#\ of\ items\ rated\ partially\ aligned\ by\ reviewer_k}{\#\ of\ items}\right)}{K}$$

Where K is the total number or raters. We will use grade 6 mathematics, reporting category 2 (from Table 4 of the results section) as an example. The reporting category includes 20 items and three reviewers provided ratings. One reviewer rated two of the 20 items as "partially aligned", the second reviewer rated one of the 20 items as "partially aligned", and the third reviewer did not rate any of the items as "partially aligned". Using the formula above the average percentage of items rated as partially aligned among the three raters is:

$$Average\ \% = \frac{\left(\frac{2}{20}+\frac{1}{20}+\frac{0}{20}\right)}{3} = \ .05\ (or\ 5\%)$$

This does not mean 5% of the items are partially aligned to the TEKS content standards. Rather this is the average *percentage* of items assigned a "partially aligned" rating among reviewers. Each reviewer may have identified the same item, or the reviewers may have identified different items. In the case of category 2 for grade 6 – two reviewers rated the same item as "partially" aligned and one reviewer rated a different item as "partially aligned". The results tables included in this report provide information about the number of reviewers per item rated "partially aligned" or "not aligned".

We used the same approach to compute the average percentage of items rated "fully aligned" and "not aligned". We conducted analyses overall and by categories identified in the blueprints – reporting category, standard type (readiness or supporting), and item type, when applicable. The results tables summarize the content review information for each grade and content area.

Results

Mathematics

The Texas mathematics assessments include four reporting categories: (a) Numerical Representations and Relationships, (b) Computations and Algebraic Relationships, (c) Geometry and Measurement, and (d) Data Analysis and Personal Finance Literacy. Mathematics includes readiness and supporting standards, and the test forms include multiple choice and gridded items.

Table 1 presents the content review results for the 2016 grade 3 mathematics STAAR test form. The number of items included on the test form matched the blueprint overall, as well as, disaggregated by reporting category, standard type, and item type.

All grade 3 mathematics items falling under reporting categories 2, 3, and 4 were rated as “fully aligned” to the intended TEKS expectation by all three reviewers. For category 1, the average percentage of items rated as “fully aligned” to the intended TEKS expectation, averaged among the three reviewers, was 91.7%. Three items were rated as “partially aligned” by one reviewer.

Table 1. Grade 3 Mathematics Content Alignment and Blueprint Consistency Results

Category	Blueprint # Questions	Form # Questions	Average Percentage of items rated Fully Aligned to Expectation among Reviewers	Average Percentage of items rated Partially Aligned to Expectation among Reviewers	Number of Items Rated as Partially Aligned by One or more Reviewer	Average Percentage of items rated Not Aligned to Expectation among Reviewers	Number of Items Rated as Not Aligned by One or more Reviewer
Reporting Category							
1: Numerical Representations and Relationships	12	12	91.7%	8.3%	Three items by one reviewer each	0.0%	--
2: Computations and Algebraic Relationships	18	18	100.0%	0.0%	--	0.0%	--
3: Geometry and Measurement	10	10	100.0%	0.0%	--	0.0%	--
4: Data Analysis and Personal Finance Literacy	6	6	100.0%	0.0%	--	0.0%	--
Standard Type							
Readiness Standards	28-30	28	96.4%	3.6%	Three items by one reviewer each	0.0%	--
Supporting Standards	16-18	18	100.0%	0.0%	--	0.0%	--
Item Type							
Multiple Choice	43	43	97.7%	2.3%	Three items by one reviewer each	0.0%	--
Gridded	3	3	100.0%	0.0%	--	0.0%	--
Total	**46**	**46**	**97.8%**	**2.2%**	**Three items**	**0.0%**	--

A summary of the content review results for the 2016 grade 4 mathematics STAAR test form is presented in Table 2. The number of items included on the test form matched the blueprint overall, as well as, disaggregated by reporting category, standard type, and item type.

All three reviewers rated all grade 4 mathematics items falling under reporting category 4 as “fully aligned” to the intended TEKS expectations. For reporting categories 1, 2 and 3, the average percentage of items rated “fully aligned” to the intended expectation, averaged among the three reviewers, were 94.4%, 97.9%, and 95.6%, respectively. Two items in reporting category 1, one item in reporting category 2, and two items in reporting category 3 were rated “partially aligned” by one reviewer.

Table 2. Grade 4 Mathematics Content Alignment and Blueprint Consistency Results

Category	Blueprint # Questions	Form # Questions	Average Percentage of items rated Fully Aligned to Expectation among Reviewers	Average Percentage of items rated Partially Aligned to Expectation among Reviewers	Number of Items Rated as Partially Aligned by One or more Reviewer	Average Percentage of items rated Not Aligned to Expectation among Reviewers	Number of Items Rated as Not Aligned by One or more Reviewer
	Reporting Category						
1: Numerical Representations and Relationships	12	12	94.4%	5.6%	Two items by one reviewer each	0.0%	--
2: Computations and Algebraic Relationships	16	16	97.9%	2.1%	One item by one reviewer	0.0%	--
3: Geometry and Measurement	15	15	95.6%	4.4%	Two items by one reviewer each	0.0%	--
4: Data Analysis and Personal Finance Literacy	5	5	100.0%	0.0%	--	0.0%	--
	Standard Type						
Readiness Standards	29-31	30	95.6%	4.4%	Four items by one reviewer each	0.0%	--
Supporting Standards	17-19	18	98.1%	1.9%	One item by one reviewer	0.0%	--
	Item Type						
Multiple Choice	45	45	97.0%	3.0%	Four items by one reviewer each	0.0%	--
Gridded	3	3	88.9%	11.1%	One item by one reviewer	0.0%	--
Total	**48**	**48**	**96.5%**	**3.5%**	**Five items**	**0.0%**	--

Table 3 presents the content review results for the 2016 grade 5 mathematics STAAR test form. The number of items included on the test form matched the blueprint overall, as well as, disaggregated by reporting category, standard type, and item type.

All grade 5 mathematics items falling under reporting categories 1, 3, and 4 were rated as “fully aligned” to the intended TEKS expectation by all four reviewers. For reporting category 2, the average percentage of items rated as “fully aligned” to the intended expectation, averaged among the four reviewers, was approximately 97%. Three items in reporting category 2 were rated as “partially aligned” by one reviewer each.

Table 3. Grade 5 Mathematics Content Alignment and Blueprint Consistency Results

Category	Blueprint # Questions	Form # Questions	Average Percentage of items rated Fully Aligned to Expectation among Reviewers	Average Percentage of items rated Partially Aligned to Expectation among Reviewers	Number of Items Rated as Partially Aligned by One or more Reviewer	Average Percentage of items rated Not Aligned to Expectation among Reviewers	Number of Items Rated as Not Aligned by One or more Reviewer
	Reporting Category						
1: Numerical Representations and Relationships	8	8	100.0%	0.0%	--	0.0%	--
2: Computations and Algebraic Relationships	24	24	96.9%	3.1%	Three items by one reviewer each	0.0%	--
3: Geometry and Measurement	12	12	100.0%	0.0%	--	0.0%	--
4: Data Analysis and Personal Finance Literacy	6	6	100.0%	0.0%	--	0.0%	--
Readiness Standards	30-33	31	98.4%	1.6%	Two items by one reviewer each	0.0%	--
Supporting Standards	17-20	19	98.7%	1.3%	One item by one reviewer	0.0%	--
Multiple Choice	47	47	98.4%	1.6%	Three items by one reviewer each	0.0%	--
Gridded	3	3	100.0%	0.0%	--	0.0%	--
Total	**50**	**50**	**98.5%**	**1.5%**	**Three items**	0.0%	**--**

The content review results for the 2016 grade 6 mathematics STAAR test form are presented in Table 4. The number of items included on the test form matched the blueprint overall, as well as, disaggregated by reporting category, standard type, and item type.

All grade 6 mathematics items falling under reporting categories 1 and 4 were rated as “fully aligned” to the intended expectation by all three reviewers. For reporting categories 2 and 3, the average percentages of items rated as “fully aligned” to the intended expectation, averaged among the three reviewers, were 95% and 95.8%, respectively. For reporting category 2, two reviewers rated one item as “partially aligned” and one reviewer rated a different item as “partially aligned”. For category 3, one reviewer rated one item as “partially aligned”.

Table 4. Grade 6 Mathematics Content Alignment and Blueprint Consistency Results

Category	Blueprint # Questions	Form # Questions	Average Percentage of items rated Fully Aligned to Expectation among Reviewers	Average Percentage of items rated Partially Aligned to Expectation among Reviewers	Number of Items Rated as Partially Aligned by One or more Reviewer	Average Percentage of items rated Not Aligned to Expectation among Reviewers	Number of Items Rated as Not Aligned by One or more Reviewer
	Reporting Category						
1: Numerical Representations and Relationships	14	14	100.0%	0.0%	--	0.0%	--
2: Computations and Algebraic Relationships	20	20	95.0%	5.0%	One item by one reviewer; One item by two reviewers	0.0%	--
3: Geometry and Measurement	8	8	95.8%	4.2%	One item by one reviewer	0.0%	--
4: Data Analysis and Personal Finance Literacy	10	10	100.0%	0.0%	--	0.0%	--
Standard Type							
Readiness Standards	31-34	33	97.0%	3.0%	One item by one reviewer; One item by two reviewers	0.0%	--
Supporting Standards	18-21	19	98.2%	1.8%	One item by one reviewer	0.0%	--
Item Type							
Multiple Choice	48	48	97.2%	2.8%	Two items by one reviewer each; One item by two reviewers	0.0%	--
Gridded	4	4	100.0%	0.0%	--	0.0%	--
Total	**52**	**52**	**97.4%**	**2.6%**	**Three items**	**0.0%**	**--**

Table 5 presents the content review results for the 2016 grade 7 mathematics STAAR test form. The number of items included on the test form matched the blueprint overall, as well as, disaggregated by reporting category, standard type, and item type.

All grade 7 mathematics items falling under reporting categories 1 and 2 were rated as “fully aligned” to the intended expectation by all three reviewers. For reporting categories 3 and 4, the average percentage of items rated “fully aligned” to the intended expectation, averaged among reviewers, were 97.9% and 96.3%, respectively. For each of these two reporting categories, one reviewer rated one item as “partially aligned” to the intended expectation.

Table 5. Grade 7 Mathematics Content Alignment and Blueprint Consistency Results

Category	Blueprint # Questions	Form # Questions	Average Percentage of items rated Fully Aligned to Expectation among Reviewers	Average Percentage of items rated Partially Aligned to Expectation among Reviewers	Number of Items Rated as Partially Aligned by One or more Reviewer	Average Percentage of items rated Not Aligned to Expectation among Reviewers	Number of Items Rated as Not Aligned by One or more Reviewer
	Reporting Category						
1: Numerical Representations and Relationships	9	9	100.0%	0.0%	--	0.0%	--
2: Computations and Algebraic Relationships	20	20	100.0%	0.0%	--	0.0%	--
3: Geometry and Measurement	16	16	97.9%	2.1%	One item by one reviewer	0.0%	--
4: Data Analysis and Personal Finance Literacy	9	9	96.3%	3.7%	One item by one reviewer	0.0%	--
Standard Type							
Readiness Standards	32-35	35	99.0%	1.0%	One item by one reviewer	0.0%	--
Supporting Standards	19-22	19	98.2%	1.8%	One item by one reviewer	0.0%	--
Item Type							
Multiple Choice	50	50	98.7%	1.3%	Two items by one reviewer each	0.0%	--
Gridded	4	4	100.0%	0.0%	--	0.0%	--
Total	**54**	**54**	**98.8%**	**1.2%**	**Two items**	**0.0%**	**--**

The content review results for the 2016 grade 8 mathematics STAAR test form are presented in Table 6. The number of items included on the test form matched the blueprint overall, as well as, disaggregated by reporting category, standard type, and item type.

All grade 8 mathematics items falling under reporting categories 1 and 4 were rated as “fully aligned” to the intended expectation by all four reviewers. For reporting categories 2 and 3, the average percentages of items “fully aligned” to the intended expectation, averaged among the four reviewers, were 97.7% and 96.3%, respectively. For reporting category 2, there was one item rated as “partially aligned” and one item rated as “not aligned” by one reviewer each. For reporting category 3, one item was rated as “partially aligned” by one reviewer and one item was rated “not aligned” by two reviewers.

Table 6. Grade 8 Mathematics Content Alignment and Blueprint Consistency Results

Category	Blueprint # Questions	Form # Questions	Average Percentage of items rated Fully Aligned to Expectation among Reviewers	Average Percentage of items rated Partially Aligned to Expectation among Reviewers	Number of Items Rated as Partially Aligned by One or more Reviewer	Average Percentage of items rated Not Aligned to Expectation among Reviewers	Number of Items Rated as Not Aligned by One or more Reviewer
	Reporting Category						
1: Numerical Representations and Relationships	5	5	100.0%	0.0%	--	0.0%	--
2: Computations and Algebraic Relationships	22	22	97.7%	1.1%	One item by one reviewer	1.1%	One item by one reviewer
3: Geometry and Measurement	20	20	96.3%	1.3%	One item by one reviewer	2.5%	One item by two reviewers
4: Data Analysis and Personal Finance Literacy	9	9	100.0%	0.0%	--	0.0%	--
Readiness Standards	34-36	36	97.9%	0.7%	One item by one reviewer	1.4%	One item by two reviewers
Supporting Standards	20-22	20	97.5%	1.3%	One item by one reviewer	1.3%	One item by one reviewer
Multiple Choice	52	52	98.1%	0.5%	One item by one reviewer	1.4%	One item by one reviewer; one item by two reviewers
Gridded	4	4	93.8%	6.3%	One item by one reviewer	0.0%	--
Total	**56**	**56**	**97.8%**	**0.9%**	**Two items**	**2.2%**	**Two items**

Reading

The Texas reading assessments include three reporting categories: (a) Understanding/Analysis across Genres, (b) Understanding/Analysis of Literary Texts, and (c) Understanding/Analysis of Informational Texts. Reading includes readiness and supporting standards. All STAAR reading assessment items are multiple choice.

Table 7 presents the content review results for the 2016 grade 3 reading STAAR test form. The number of items included on the test form matched the blueprint overall, as well as at each of the three reporting categories, and for each standard type.

The average percentage of grade 3 reading items rated “fully aligned” to the intended expectation, when averaged among the four reviewers, was 86.2%. For reporting categories 1, 2, and 3, these percentages were 95.8%, 94.4%, and 75%, respectively. Reporting category 3, includes one constructed response item, which was rated as “partially aligned” by one reviewer. Across all reporting categories, there were 16 items with at least one “partially aligned” rating among the four reviewers, and two items with one rating of “not aligned”.

Table 7. Grade 3 Reading Content Alignment and Blueprint Consistency Results

Category	Blueprint # Questions	Form # Questions	Average Percentage of items rated Fully Aligned to Expectation among Reviewers	Average Percentage of items rated Partially Aligned to Expectation among Reviewers	Number of Items Rated as Partially Aligned by One or more Reviewer	Average Percentage of items rated Not Aligned to Expectation among Reviewers	Number of Items Rated as Not Aligned by One or more Reviewer
	Reporting Category						
1: Understanding/ Analysis across Genres	6	6	95.8%	4.2%	One item by one reviewer	0.0%	--
2: Understanding/ Analysis of Literary Texts	18	18	94.4%	5.6%	Four items by one reviewer each	0.0%	--
3: Understanding/ Analysis of Informational Texts	16	16	73.4%	23.4%	One item by three reviewers; two items by two reviewers each; eight items by one reviewer each	3.1%	Two items by one reviewer each
Readiness Standards	24-28	25	81.0%	17.0%	One item by three reviewers; two items by two reviewers each; ten items by one reviewer each	2.0%	Two items by one reviewer each
Supporting Standards	12-16	15	95.0%	5.0%	Three items by one reviewer each	0.0%	--
Total	**40**	**40**	**86.2%**	**12.5%**	**16 items**	**1.2%**	**Two items**

The content review results for the 2016 grade 4 reading STAAR test form are presented in Table 8. The number of items included on the test form matched the blueprint overall, as well as and when disaggregated by reporting category and standard type.

The average percentage of grade 4 reading items rated as “fully aligned” to the intended expectation, averaged among the four reviewers, was 91.5%. For reporting category 1, all items were rated as “fully aligned” by all reviewers. For reporting category 2, at least one reviewer assigned a rating of “partially aligned” to six items and one reviewer rated one item as “not aligned”. For items falling under reporting category 3, there were four items rated as “partially aligned” by one reviewer each, and one item rated as “not aligned” by one reviewer.

Table 8. Grade 4 Reading Content Alignment and Blueprint Consistency Results

Category	Blueprint # Questions	Form # Questions	Average Percentage of items rated Fully Aligned to Expectation among Reviewers	Average Percentage of items rated Partially Aligned to Expectation among Reviewers	Number of Items Rated as Partially Aligned by One or more Reviewer	Average Percentage of items rated Not Aligned to Expectation among Reviewers	Number of Items Rated as Not Aligned by One or more Reviewer
	Reporting Category						
1: Understanding/ Analysis across Genres	10	10	100.0%	0.0%	--	0.0%	--
2: Understanding/ Analysis of Literary Texts	18	18	90.3%	8.3%	Six items by one reviewer each	1.4%	One item by one reviewer
3: Understanding/ Analysis of Informational Texts	16	16	87.5%	10.9%	One item by three reviewers; one item by two reviewers; Two items by one reviewer each	1.6%	One item by one reviewer
Readiness Standards	26-31	29	89.7%	8.6%	One item by three reviewers; one item by two reviewers; five items by one reviewer each	1.7%	Two items by one reviewer each
Supporting Standards	13-18	15	95.0%	5.0%	Three items by one reviewer each	0.0%	--
Total	**44**	**44**	**91.5%**	**7.4%**	**10 items**	**1.2%**	**Two items**

Table 9 presents the content review results for the 2016 grade 5 reading STAAR test form. The number of items included on the test form matched the blueprint overall, as well as at each of the three reporting categories, and for each standard type.

Overall and for all reporting categories, the majority of items were rated as “fully aligned” to the expectation for grade 5 reading. For reporting categories 1, 2 and 3, the average percentage of items rated “fully aligned” to the intended expectation, averaged among the four reviewers, were 95%, 88.2%, and 85.3%, respectively. One item in reporting category 1, six items in reporting category 2, and six items in category 3 were rated as “partially aligned” by at least one reviewer. One item in category 1, three items in category 2, and one item in category 3 were rated as “not aligned” by one reviewer.

Table 9. Grade 5 Reading Content Alignment and Blueprint Consistency Results

Category	Blueprint # Questions	Form # Questions	Average Percentage of items rated Fully Aligned to Expectation among Reviewers	Average Percentage of items rated Partially Aligned to Expectation among Reviewers	Number of Items Rated as Partially Aligned by One or more Reviewer	Average Percentage of items rated Not Aligned to Expectation among Reviewers	Number of Items Rated as Not Aligned by One or more Reviewer
	Reporting Category						
1: Understanding/ Analysis across Genres	10	10	95.0%	2.5%	One item by one reviewer	2.5%	One item by one reviewer
2: Understanding/ Analysis of Literary Texts	19	19	88.2%	7.9%	Six items by one reviewer each	3.9%	Three items by one reviewer each
3: Understanding/ Analysis of Informational Texts	17	17	85.3%	13.2%	Three items by two reviewers each; Three items by one reviewer each	1.5%	One item by one reviewer
Readiness Standards	28-32	29	90.5%	6.9%	Two items by two reviewers each; four items by one reviewer each	2.6%	Three items by one reviewer each
Supporting Standards	14-18	17	85.3%	11.8%	One item by two reviewers; six items by one reviewer each	2.9%	Two items by one reviewer each
Total	**46**	**46**	**88.6%**	**8.7%**	**13 items**	**2.7%**	**Five items**

Table 10 presents the content review results for the 2016 grade 6 reading STAAR test form. The number of items included on the test form matched the blueprint overall, as well as at each of the three reporting categories, and for each standard type.

Overall, the average percentage of items rated as “fully aligned” to the intended expectation, averaged among the four reviewers, was 95.8% for grade 6 reading. Broken down by reporting category these percentages were 100%, 95.5%, and 94.4% for categories 1, 2, and 3, respectively. There were seven items overall with at least one reviewer providing a rating of “partially aligned” and no items were rated as “not aligned”.

Table 10. Grade 6 Reading Content Alignment and Blueprint Consistency Results

Category	Blueprint # Questions	Form # Questions	Average Percentage of items rated Fully Aligned to Expectation among Reviewers	Average Percentage of items rated Partially Aligned to Expectation among Reviewers	Number of Items Rated as Partially Aligned by One or more Reviewer	Average Percentage of items rated Not Aligned to Expectation among Reviewers	Number of Items Rated as Not Aligned by One or more Reviewer
	Reporting Category						
1: Understanding/ Analysis across Genres	10	10	100.0%	0.0%	--	0.0%	--
2: Understanding/ Analysis of Literary Texts	20	20	95.5%	5.0%	Four items by one reviewer each	0.0%	--
3: Understanding/ Analysis of Informational Texts	18	18	94.4%	5.6%	One item by two reviewers; two items by one reviewer each	0.0%	--
Readiness Standards	29-34	31	96.8%	3.2%	Four items by one reviewer each	0.0%	--
Supporting Standards	14-19	17	94.1%	5.9%	One item by two reviewers; two items by one reviewer each	0.0%	--
Total	**48**	**48**	**95.8%**	**4.2%**	**Seven items**	**0.0%**	**--**

Table 11 presents the content review results for the 2016 grade 7 reading STAAR test form. The number of items included on the test form matched the blueprint overall, for each of the three reporting categories, and for each standard type.

For reporting categories 1, 2, and 3, the average percentage of items rated “fully aligned” to the intended expectation, averaged among the four reviewers, were 95%, 97.6%, and 80.3%, respectively. One item in category 1, two items in category 2, and seven items in category 3 were rated as “partially aligned” by one or more reviewers. One reviewer rated one item in reporting category 3 as “not aligned”.

Table 11. Grade 7 Reading Content Alignment and Blueprint Consistency Results

Category	Blueprint # Questions	Form # Questions	Average Percentage of items rated Fully Aligned to Expectation among Reviewers	Average Percentage of items rated Partially Aligned to Expectation among Reviewers	Number of Items Rated as Partially Aligned by One or more Reviewer	Average Percentage of items rated Not Aligned to Expectation among Reviewers	Number of Items Rated as Not Aligned by One or more Reviewer
	Reporting Category						
1: Understanding/ Analysis across Genres	10	10	95.0%	5.0%	One item by two reviewers	0.0%	--
2: Understanding/ Analysis of Literary Texts	21	21	97.6%	2.4%	Two items by one reviewer each	0.0%	--
3: Understanding/ Analysis of Informational Texts	19	19	80.3%	18.4%	Three items by three reviewers each; one item by two reviewers; Three items by one reviewer each	1.3%	One item by one reviewer
Readiness Standards	30-35	31	87.9%	11.3%	Three items by three reviewers each; two items by two reviewers each; one item by one reviewer	0.8%	One item by one reviewer
Supporting Standards	15-20	19	94.8%	5.2%	Four items by one reviewer	0.0%	--
Total	**50**	**50**	**90.5%**	**9.0%**	**Ten items**	**0.5%**	**One item**

The content review results for the 2016 grade 8 reading STAAR test form are presented in Table 12. The number of items included on the test form matched the blueprint overall, as well as when disaggregated by reporting category and standard type.

All grade 8 reading items falling under reporting category 1 were rated as “fully aligned” to the intended expectations by all four reviewers. For reporting categories 1 and 2, the average percentage of items rated “fully aligned” to the intended expectation, averaged among the three reviewers, were 96.6% and 95.0%, respectively. Three items in reporting category 2 were rated as “partially aligned” by one reviewer each, and one item in reporting category 3 was rated as “partially aligned” by two reviewers. One item in reporting category 3 was rated “not aligned” by two reviewers.

Table 12. Grade 8 Reading Content Alignment and Blueprint Consistency Results

Category	Blueprint # Questions	Form # Questions	Average Percentage of items rated Fully Aligned to Expectation among Reviewers	Average Percentage of items rated Partially Aligned to Expectation among Reviewers	Number of Items Rated as Partially Aligned by One or more Reviewer	Average Percentage of items rated Not Aligned to Expectation among Reviewers	Number of Items Rated as Not Aligned by One or more Reviewer
	Reporting Category						
1: Understanding/ Analysis across Genres	10	10	100.0%	0.0%	--	0.0%	--
2: Understanding/ Analysis of Literary Texts	22	22	96.6%	3.4%	Three items by one reviewer each	0.0%	--
3: Understanding/ Analysis of Informational Texts	20	20	95.0%	2.5%	One item by two reviewers	2.5%	One item by two reviewers
Readiness Standards	31-36	32	96.9%	3.1%	One item by two reviewers; two items by one reviewer each	0.0%	--
Supporting Standards	16-21	20	96.3%	1.3%	One item by one reviewer	2.5%	One item by two reviewers
Total	**52**	**52**	**96.6%**	**2.4%**	**Four items**	**1.0%**	**One item**

Science

The Texas science assessments include four reporting categories: (a) Matter and Energy, (b) Force, Motion, and Energy, (c) Earth and Space, and (d) Organisms and Environments. Science includes readiness and supporting standards. The STAAR science assessments include primarily multiple choice with a small number of gridded items.

Table 13 presents the content review results for the 2016 grade 5 science STAAR test form. The number of items included on the test form matched the blueprint overall, as well as when disaggregated by reporting category, standard type, and item type.

The average percentage of grade 5 science items rated “fully aligned” to the intended expectation averaged among the four reviewers, was 98.3%. All of the items falling under category 2 were rated as “fully aligned” to the intended expectations, and only one item each for reporting categories 1, 3, and 4 was rated as “partially aligned” or “not aligned” by one reviewer.

Table 13. Grade 5 Science Content Alignment and Blueprint Consistency Results

Category	Blueprint # Questions	Form # Questions	Average Percentage of items rated Fully Aligned to Expectation among Reviewers	Average Percentage of items rated Partially Aligned to Expectation among Reviewers	Number of Items Rated as Partially Aligned by One or more Reviewer	Average Percentage of items rated Not Aligned to Expectation among Reviewers	Number of Items Rated as Not Aligned by One or more Reviewer
	Reporting Category						
1: Matter and Energy	8	8	96.9%	0.0%	--	3.1%	One item by one reviewer
2: Force, Motion, and Energy	10	10	100.0%	0.0%	--	0.0%	--
3: Earth and Space	12	12	97.9%	2.1%	One item by one reviewer	0.0%	--
4: Organisms and Environments	14	14	98.2%	1.8%	One item by one reviewer	0.0%	--
Readiness Standards	26-29	28	98.2%	0.9%	One item by one reviewer	0.9%	One item by one reviewer
Supporting Standards	15-18	16	98.4%	1.6%	One item by one reviewer	0.0%	--
Multiple Choice	43	43	98.3%	1.2%	Two items by one reviewer each	0.6%	One item by one reviewer
Gridded	1	1	100.0%	0.0%	--	0.0%	--
Total	**44**	**44**	**98.3%**	**1.1%**	**Two items**	**0.6%**	**One item**

Table 14 presents the content review results for the 2016 grade 8 science STAAR test form. The number of items included on the test form matched the blueprint overall, as well as when disaggregated by reporting category, standard type, and item type.

All grade 8 science items falling under reporting categories 1 and 3 were rated as “fully aligned” to the intended TEKS expectations by all four reviewers. For reporting categories 2 and 4, the average percentage of items rated “fully aligned” to the intended expectation averaged among the three reviewers were 91.7% and 98.2%, respectively. Four items in reporting category 2 and one item in reporting category 4 were rated by one reviewer as “not aligned”.

Table 14. Grade 8 Science Content Alignment and Blueprint Consistency Results

Category	Blueprint # Questions	Form # Questions	Average Percentage of items rated Fully Aligned to Expectation among Reviewers	Average Percentage of items rated Partially Aligned to Expectation among Reviewers	Number of Items Rated as Partially Aligned by One or more Reviewer	Average Percentage of items rated Not Aligned to Expectation among Reviewers	Number of Items Rated as Not Aligned by One or more Reviewer
	Reporting Category						
1: Matter and Energy	14	14	100.0%	0.0%	--	0.0%	--
2: Force, Motion, and Energy	12	12	91.7%	0.0%	--	8.3%	Four items by one reviewer each
3: Earth and Space	14	14	100.0%	0.0%	--	0.0%	--
4: Organisms and Environments	14	14	98.2%	0.0%	--	1.8%	One item by one reviewer
Standard Type							
Readiness Standards	32-35	34	97.1%	0.0%	--	2.9%	Four items by one reviewer each
Supporting Standards	19-22	20	98.8%	0.0%	--	1.3%	One item by one reviewer
Item Type							
Multiple Choice	50	50	98.0%	0.0%	--	2.0%	Four items by one reviewer each
Gridded	4	4	93.8%	0.0%	--	6.3%	One item by one reviewer
Total	**54**	**54**	**97.7%**	**0.0%**	**--**	**2.3%**	**Five items**

Social Studies

The Texas social studies assessment, given at grade 8 only, includes four reporting categories: (a) History, (b) Geography and Culture, (c) Government and Citizenship, and (d) Economics, Science, Technology, and Society. Social studies includes readiness and supporting standards. The STAAR social studies assessment is composed of all multiple choice items.

Table 15 presents the content review results for the 2016 grade 8 social studies STAAR test form. The number of items included on the test form matched the blueprint overall, as well as when disaggregated by reporting category, standard type, and item type.

For social studies, the average percentage of items rated “fully aligned” to the intended expectation, averaged among the four reviewers, was 89.9% overall. When broken down by reporting categories 1, 2, 3, and 4, the percentage of items rated as “fully aligned” were 90%, 91.7%, 87.5%, and 90.6%, respectively. There were 13 total items across all categories rated as “partially aligned” by one or more reviewers, and three items rated as “not aligned” by at least one reviewer.

Table 15. Grade 8 Social Studies Content Alignment and Blueprint Consistency Results

Category	Blueprint # Questions	Form # Questions	Average Percentage of items rated Fully Aligned to Expectation among Reviewers	Average Percentage of items rated Partially Aligned to Expectation among Reviewers	Number of Items Rated as Partially Aligned by One or more Reviewer	Average Percentage of items rated Not Aligned to Expectation among Reviewers	Number of Items Rated as Not Aligned by One or more Reviewer
	Reporting Category						
1: History	20	20	90.0%	6.3%	One item by two reviewers; three items by one reviewer each	3.8%	One item by two reviewers; one item by one reviewer
2: Geography and Culture	12	12	91.7%	8.3%	One item by two reviewers; two items by one reviewer each	0.0%	--
3: Government and Citizenship	12	12	87.5%	8.3%	One item by two reviewers; two items by one reviewer each	4.2%	One item by two reviewers
4: Economics, Science, Technology, and Society	8	8	90.6%	9.4%	Three items by one reviewer each	0.0%	--
Readiness Standards	31-34	34	89.0%	8.8%	Two items by two reviewers each; seven items by one reviewer each	2.2%	One item by two reviewers; one item by one reviewer
Supporting Standards	18-21	18	91.7%	5.6%	Four items by one reviewer each	2.8%	One item by two reviewers
Total	**52**	**52**	**89.9%**	**7.7%**	**13 items**	**2.4%**	**Three items**

Writing

The Texas writing assessments include three reporting categories: (a) Composition, (b) Revision, and (c) Editing. Writing includes readiness and supporting standards. STAAR writing assessments include one composition item, and the remaining items are multiple choice.

Table 16 presents content review results for the 2016 grade 4 writing STAAR test form. The number of items included on the test form matched the blueprint overall, as well as when disaggregated by reporting category, standard type, and item type.

All four reviewers rated all grade 4 writing items falling under reporting category 2 as “fully aligned” to the intended expectations. For reporting categories 1 and 3, the average percentage of items rated “fully aligned” to the intended expectation, averaged among the three reviewers, were 75% and 91.7%, respectively. One item in reporting category 1 and three items in reporting category 3 were rated by one reviewer as “partially aligned”. One reviewer rated one item as “not aligned”.

Table 16. Grade 4 Writing Content Alignment and Blueprint Consistency Results

Category	Blueprint # Questions	Form # Questions	Average Percentage of items rated Fully Aligned to Expectation among Reviewers	Average Percentage of items rated Partially Aligned to Expectation among Reviewers	Number of Items Rated as Partially Aligned by One or more Reviewer	Average Percentage of items rated Not Aligned to Expectation among Reviewers	Number of Items Rated as Not Aligned by One or more Reviewer
	Reporting Category						
1: Composition	1	1	75.0%	25.0%	One item by one reviewer	0.0%	--
2: Revision	6	6	100.0%	0.0%	--	0.0%	--
3: Editing	12	12	91.7%	6.3%	Three items by one reviewer each	2.1%	One item by one reviewer
Readiness Standards	11-13	14	94.6%	5.4%	Three items by one reviewer each	0.0%	--
Supporting Standards	5-7	5	90.0%	5.0%	One item by one reviewer	5.0%	One item by one reviewer
Multiple Choice	18	18	94.5%	4.2%	Three items by one reviewer each	1.4%	One item by one reviewer
Composition	1	1	75.0%	25.0%	One item by one reviewer	0.0%	--
Total	**19**	**19**	**93.4%**	**5.3%**	**Four items**	**1.3%**	**One item**

The 2016 grade 7 writing STAAR test form content review results are presented in Table 17. The number of items included on the test form matched the blueprint overall, as well as at each reporting category, for each standard type, and by item type.

For reporting categories 1, 2 and 3, the average percentage of items rated fully aligned to the intended expectation, averaged among the four reviewers, were 75%, 84.6% and 92.6%, respectively. Across the entire form, there were eight items rated as “partially aligned” and four items rated “not aligned” by at least one reviewer.

Table 17. Grade 7 Writing Content Alignment and Blueprint Consistency Results

Category	Blueprint # Questions	Form # Questions	Average Percentage of items rated Fully Aligned to Expectation among Reviewers	Average Percentage of items rated Partially Aligned to Expectation among Reviewers	Number of Items Rated as Partially Aligned by One or more Reviewer	Average Percentage of items rated Not Aligned to Expectation among Reviewers	Number of Items Rated as Not Aligned by One or more Reviewer
	Reporting Category						
1: Composition	1	1	75.0%	25.0%	One item by one reviewer	0.0%	--
2: Revision	13	13	84.6%	5.8%	Three items by one reviewer each	9.6%	Two items by two reviewers each; one item by one reviewer
3: Editing	17	17	92.6%	5.9%	Four items by one reviewer each	1.5%	One item by one reviewer
Readiness Standards	18-21	20	91.3%	6.3%	Five items by one reviewer each	2.5%	Two items by one reviewer each
Supporting Standards	9-12	11	84.1%	6.8%	Three items by one reviewer each	9.1%	Two items by two reviewers each
Multiple Choice	30	30	89.1%	5.9%	Seven items by one reviewer each	5.0%	Two items by two reviewers each; two items by one reviewer each
Composition	1	1	75.0%	25.0%	One item by one reviewer	0.0%	--
Total	**31**	**31**	**88.7%**	**6.5%**	**Eight items**	**4.8%**	**Four items**

Content Review Summary and Discussion

HumRRO's content review provided evidence to support the content validity of the 2016 STAAR test forms for mathematics and reading grades 3 through 8, science grades 5 and 8, social studies grade 8, and writing grades 4 and 7. Overall, the test forms were found to be consistent with the blueprints and TEKS documentation.

The numbers of items included on the assessment forms were consistent with the blueprint for all grades and content areas reviewed. Additionally, the results provide evidence that the 2016 STAAR test forms are well-aligned to the intended TEKS expectations. This was true at the total assessment form level and when examining results by reporting category, standards type, and item-type. Mathematics had a particularly high average percentage of items rated as fully aligned. Grade 7 writing included the highest percentage of items rated as not aligned; however, this represented fewer than five percent of the overall items, and the majority of items rated 'not aligned' to the intended TEKS expectation were rated as aligning to a different TEKS student expectation within the same reporting category.

Task 2: Replication and Estimation of Reliability and Measurement Error

Estimation of Reliability and Measurement Error

Internal consistency reliability and standard error of measurement (SEM) estimates cannot be computed for a test until student response data are available. However, we can make projections about the reliability and SEM using the: (a) IRT parameter estimates that were used to construct test forms and (b) projections of the distribution of student scores. We used the Kolen, Zang, and Hanson (1996; KZH) procedures to compute internal consistency reliability estimates as well as overall and conditional SEMs.

For reading and mathematics, the number of items on each assessment was consistent for 2015 and 2016. We used the 2015 student cumulative frequency distribution (CFD) for STAAR scores as the projected 2016 distribution. For writing, where the test form was shorter for 2016, we interpolated the 2015 STAAR score CFD onto the shorter 2016 scale to find the projected 2016 raw score mean and standard deviation. We smoothed the CFD by computing a normal distribution with the projected mean and standard deviation.

The projected internal consistency reliability and overall SEM estimates for mathematics and reading grades 3 through 8, science grades 5 and 8, social studies grade 8, and writing grades 4 and 7 are presented in Table 18. Internal consistency reliability estimates are measures of the relationship among items that are purported to measure a common construct. Overall, the reliability estimates are acceptable to excellent. Internal consistency estimates above 0.70 are typically considered acceptable, with estimates of 0.90 and higher considered excellent (Nunnally, 1978). The projected SEM provides an estimate of how close students' observed scores are to their true scores. For example, on average, for reading grade 5, students' observed STAAR scores are projected to be plus or minus 2.75 raw score points from their true score. Appendix A provides figures of the CSEMs across the raw STAAR score distribution. CSEM plots tend to be U-shaped, with lower SEMs in the center of the distribution and higher SEMs at the lower and upper ends of the distribution. These results are reasonable and typical of most testing programs.

There are a number of factors that contribute to reliability estimates, including test length and item types. Typically, longer tests tend to have higher reliability and lower SEMs. Additionally, mixing item types such as multiple choice items and composition items may result in lower reliability estimates. The lower reliability estimates for writing are not surprising, given there are two item types and fewer items overall, especially for grade 4. Most testing programs accept lower reliability estimates for writing tests because they recognize that composition items are able to measure an aspect of the writing construct that multiple choice items cannot. This combination of different item formats can increase the content evidence for the validity of test scores, which is more important than the slight reduction in reliability.

Overall, the projected reliability and SEM estimates are reasonable.

Table 18. Projected Reliability and SEM Estimates

Subject	Grade	KZH Projected Reliability	KZH Projected SEM
Mathematics	3	0.918	2.77
Mathematics	5	0.913	3.09
Mathematics	4	0.916	2.80
Mathematics	6	0.925	3.09
Mathematics	7	0.922	3.10
Mathematics	8	0.907	3.14
Reading	3	0.890	2.65
Reading	4	0.913	2.71
Reading	5	0.908	2.75
Reading	6	0.910	2.84
Reading	7	0.903	2.96
Reading	8	0.914	2.94
Science	5	0.883	2.74
Science	8	0.906	3.05
Social Studies	8	0.895	3.19
Writing	4	0.786	1.99
Writing	7	0.846	3.10

Replication of Calibration and Equating Procedures

We conducted a procedural replication of the 2015 calibration and equating process. Following the 2015 STAAR equating specifications (made available to HumRRO), we conducted calibration analyses on the 2015 operational items for mathematics, reading, social studies, science and writing. For reading, science, social studies, and writing, we also conducted equating analyses to put the 2015 operational items onto the STAAR’s scale. Finally, we calibrated and equated the field test items for all grades and subjects. Overall, the procedures used by the primary contractor to calibrate and equate operational and field test items are acceptable and should result in test scores for a given grade having the same meaning year to year.

We are concerned that no composition items were included in the equating item set for writing. As noted in the STAAR equating specifications document, it is important to examine the final equating set for content representation. The equating set should represent the continuum of the content tested. By excluding composition items from the equating set, Texas is limited in being able to adjust for year-to-year differences in content that is covered by the composition items. However, this is not an uncommon practice for large-scale testing programs. There are many practical limitations to including open-response items in the equating set. Notably, typically only one or two open-response items are included on an exam and this type of item tends to be very memorable. Including open-response items in the equating set requires repeating the item year to year, increasing the likelihood of exposure. The risk of exposure typically outweighs the benefit of including the item type in the equating set.

Task 3: Judgments about Validity and Reliability based on Review of STAAR Documentation

Background

While Tasks 1 and 2 were devoted to empirical evidence, this section reports HumRRO's subjective judgements about the validity and reliability for 2016 STAAR scores based on a review of the processes used to build and administer the assessments. There are two important points in this lead statement.

First, certain types of evidence for validity and reliability can only be gathered after tests are administered and scores computed. However, score validity and reliability depend on the quality of all of the processes used to produce student test scores. In this section, the focus is on the *potentia*l for acceptable validity and reliability for the 2016 STAAR forms, given the procedures used to build and score the tests. Fortunately, student achievement testing is built on a long history of discovering and generating processes that create validity and reliability of assessment scores. Thus, Task 3 focuses on judgments of the processes used to produce the 2016 suite of assessments.

Second, the veracity of such judgments is based on the expertise and experience of those making the judgments. HumRRO believes that we were invited to conduct this review because of the unique role that our staff have played over the last 20 years in the arena of state- and national-level student achievement testing. HumRRO has become nationally known for its services as a quality-assurance vendor conducting research studies and replicating psychometric processes.

HumRRO began building a reputation for sound, impartial work for state assessments in 1996 when it acquired its first contract with the Department of Education for the Commonwealth of Kentucky. Over the course of twenty years, we have conducted psychometric studies and analyses for California, Florida, Utah, Minnesota, North Dakota, Pennsylvania, Massachusetts, Oklahoma, Nevada, Indiana, New York, the National Assessment of Education Progress (NAEP) and the Partnership for Assessment of Readiness for College and Careers (PARCC) assessment consortium. HumRRO also conducted an intensive one-time review of the validity and reliability of Idaho's assessment system. Additionally, HumRRO staff began conducting item content reviews for the National Research Council in the late 1990s with the Voluntary National Test initiative, followed by item reviews for California's high school exit exam. Since then HumRRO has conducted alignment studies for California, Missouri, Florida, Minnesota, Kentucky, Colorado, Tennessee, Georgia, the National Assessment Governing Board (NAGB) and the Smarter Balance assessment consortium.

We indicated above that HumRRO has played a unique role in assessment. We are not, however, a "major testing company" in the state testing arena in the sense that HumRRO has neither written test items nor constructed test forms for state assessments.[8] Thus, for each of the state assessments that we have been involved with, HumRRO has been required to work with that state's prime test vendor. The list of such vendors includes essentially all of the major

[8] We are, however, a full service testing company in other arenas such as credentialing and tests for hiring and promoting within organizations. Efforts in these areas include writing items, constructing forms, scoring, and overseeing test administration.

state testing contractors.[9] As a result, we have become very familiar with the processes used by the major vendors in educational testing.

Thus, the HumRRO staff assigned to Task 3 provides Texas with an excellent technical and practical foundation from which to judge the strengths and weakness of the processes for creating validity and reliability for STAAR scores. Note that while our technical expertise and experience will be used to structure our conclusions, the intent of this report is to present those conclusions so that they are accessible to a wide audience.

Basic Score Building Processes

We began our delineation of the processes we reviewed by first noting that because our focus is on test scores and test score interpretations, our review considers the processes used to create, administer, and score STAAR. The focus of our review is not on tests *per se*, but on test scores and test score uses. There are a number of important processes that must occur between *having a test* and *having a test score that is valid for a particular purpose*.

Briefly, we examined documentation of the following processes, clustered into the five major categories that lead to meaningful STAAR on-grade scores, which are to be used to compare knowledge and skill achievements of students for a given grade/subject.

1. Identify test content
    - 1.1. Determine the curriculum domain via content standards
    - 1.2. Refine the curriculum domain to a testable domain and identify reportable categories from the content standards
    - 1.3. Create test blueprints defining percentages of items for each reportable category for the test domain
2. Prepare test items
    - 2.1. Write items
    - 2.2. Conduct expert item reviews for content, bias, and sensitivity
    - 2.3. Conduct item field tests and statistical item analyses
3. Construct test forms
    - 3.1. Build content coverage into test forms
    - 3.2. Build reliability expectations into test forms
4. Administer Tests
5. Create test scores
    - 5.1. Conduct statistical item reviews for operational items
    - 5.2. Equate to synchronize scores across year
    - 5.3. Produce STAAR scores
    - 5.4. Produce test form reliability statistics

[9] At times our contracts have been directly with the state, and at other times they have been through the prime contractor as a subcontract stipulated by the state. In all cases, we have treated the state as our primary client.

Each of these processes was evaluated for its strengths in achieving on-grade student scores, which is intended to represent what a student knows and can do for a specific grade and subject. Our review was based on:

- The 2014-2015 Technical Digest, primarily Chapters 2, 3, and 4[10]
- Standard Setting Technical Report, March 15, 2013[11]
- 2015 Chapter 13 Math Standard Setting Report[12]

These documents contained references to other on-line documentation, which we also reviewed when relevant to the topics of validity and reliability. Additionally, when we could not find documentation for a specific topic area on-line, we discussed the topic with TEA and they either provided HumRRO with documents not posted on the TEA website or they described the process used for the particular topic area. Documents not posted on TEA website include the 2015 STAAR Analysis Specifications, the 2015 Standard IDM (incomplete data matrix) Analysis Specifications, and the guidelines used for test constructions. These documents expand upon the procedures documented in the Technical Digest and provided specific details that are used by all analyst to ensure consistency in results.

1. Identify Test Content

The STAAR grade/subject tests are intended to measure the critical knowledge and skills specific for a grade and subject. The validity evidence associated with the extent to which assessment scores represent students' understanding of the critical knowledge and skills starts with a clear specifications of what content should be tested. This is a three-part process that includes determining content standards, deciding which of these standards should be tested and, finally, determining what proportion of the test should cover each testable standard.

1.1. Determine content standards.

Content standards provide the foundation for score meaning by clearly and completely defining the knowledge and skills that students are to obtain for each grade/subject. For much of the history of statewide testing, grade level content standards were essentially created independently for each grade. While we have known of states adjusting their standards to connect topics from one grade to another, Texas, from the outset, took the position that content standards should flow in a logical manner from one grade to the next. That is, content for any given grade is not just important by itself. Rather it is also important in terms of how it prepares students to learn content standards for the following grade. Thus, Texas began by identifying end-of-course (EOC) objectives that support college and career readiness. From there, prerequisite knowledge and skills were determined grade by grade down to grade 3 for each of the STAAR subjects. TEA's approach to determining content standards was very thoughtful and ensures that content taught and covered in one grade links to the next grade. TEA's content standards are defined as Texas Essential Knowledge and Skills (TEKS).[13] It is beyond the

[10] http://tea.texas.gov/Student_Testing_and_Accountability/Testing/Student_Assessment_Overview/Technical_Digest_2014-2015/

[11] http://www.tea.texas.gov/WorkArea/linkit.aspx?LinkIdentifier=id&ItemID=25769804117&libID=25769804117

[12] http://www.tea.texas.gov/WorkArea/linkit.aspx?LinkIdentifier=id&ItemID=25769823236&libID=25769823334

[13] http://tea.texas.gov/curriculum/teks/

scope of this review to assess the content standards specifically. Overall, the content standards are well laid out and provide sufficient detail of the knowledge and skills that are intended to be tested by the STAAR program.

1.2. Refine testable domain.

The testable domain is a distillation of the complete TEKS domain into TEA's assessed curriculum.[14] That distillation was accomplished through "educator committee recommendations" per page 6 of the *Standard Setting Technical Report.* During this process, TEA provided guidance to committees for determining eligible and ineligible knowledge and skills. The educator committees: (a) determined the reporting categories for the assessed curriculum, (b) sorted TEKS into those reporting categories, and (c) decided which TEKS to omit from the testable domain.

1.3 Create test blueprints.

The test blueprints indicate the number, or range, of assessment items per form that should address each reporting category, standard type and item type, when applicable. The percentage of items on the blueprint representing each standard type were essentially mirrored from the assessed curriculum (70%/30% in the assessed curriculum and 65%/35% in the test blueprints, for readiness and supporting standards, respectively). The percentages of items representing each reporting category were determined through discussion with educator committees.[15]

The content standards, the assessed curriculum, and the test blueprints provide information about the knowledge and skills on which students should be tested. These materials serve as the foundation for building a test and provide the criteria by which to judge the validity of test scores.

2. Prepare Test Items

Once the testable content is defined, the test blueprints are used to guide the item writing process. This helps ensure the items measure testable knowledge and skills.

2.1. Write items.

Chapter 2 of the Technical Digest[16] provides a high-level overview of the item writing process. As described in the Technical Digest, item writers included individuals with item writing experience who are knowledgeable with specific grade content and curriculum development. Item writers are provided guidelines and are trained on how to translate the TEKS standards into items. Certainly, there is a degree of "art" or "craft" to the process of writing quality items that is difficult to fully describe in summary documents. However, overall the item writing procedures should support the development of items that measure testable content.

[14] http://tea.texas.gov/student.assessment/staar/#G_Assessments

[15] TEA provided information about this process to HumRRO during a teleconference on March 17, 2016.

[16] http://tea.texas.gov/Student_Testing_and_Accountability/Testing/Student_Assessment_Overview/Technical_Digest_2014-2015/

2.2. Conduct expert item reviews.

Chapter 2 of the Technical Digest also describes the item review process. As described in this document, items are first reviewed by the primary contractor for “the alignment between the items and the reporting categories, range of difficulty, clarity, accuracy of correct answers, and plausibility of incorrect answer choices (pg. 19).” Next, TEA staff “scrutinize each item to verify alignment to a particular student expectation in the TEKS; grade appropriateness; clarity of wording; content accuracy; plausibility of the distractors; and identification of any potential economic, regional, cultural, gender, or ethnic bias (pg. 19).” Finally, committees of Texas classroom teachers “judge each item for appropriateness, adequacy of student preparation, and any potential bias…and recommend whether the item should be field-tested as written, revised, recoded to a different eligible TEKS student expectation, or rejected (pg. 20).” The judgments, made about the alignment of each item to the TEKS expectations, provide the primary evidence that STAAR scores can be interpreted as representing students’ knowledge and skills.

2.3. Field test.

Once items have passed the hurdles described above, they are placed on operational test forms for field testing. While these field-test items are not used to produce test scores, having them intermingled among operationally scored items created the same test administration conditions (e.g., student motivation) as if they were operational items. The Technical Digest describes statistical item analyses used to show that students are responding to each individual field test item with a statistical pattern that supports the notion that higher achieving students, based on their operational test scores, tend to score higher on individual field test items and lower achieving students tend to score lower. This type of statistical analyses supports validity evidence about whether or not an item appropriately discriminates differences in grade/subject achievement. In addition, field-test statistics indicate whether or not the difficulty of the item is within the range of students’ achievement (i.e., that an individual item is neither too hard nor too easy). Item difficulty, along with item discrimination, supports both test score reliability and validity in the sense of the item contributing to measurement certainty. Note that typical item statistics cannot verify the specific reporting category or expectation-level of an item nor are they intended to do so.

Additionally, after field testing, the primary contractor and TEA curriculum and assessment specialists discuss each field test item and the associated data. Each item is reviewed for appropriateness, level of difficulty, potential bias, and reporting category/student expectation match. Based on this review, a recommendation is made on whether to accept or reject the field test item.

3. Construct Test Forms

Test form construction is critical for ensuring the items that are ultimately administered to students cover the breadth of the content that is defined as testable within the blueprint specifications. Forms are typically constructed to ensure coverage of testable content and to optimize the number of items included with high levels of discrimination that span across the ability range. The former supports validity evidence for scores, while the latter supports reliability evidence.

3.1. Build content coverage into test forms.

The blueprint provides a count of the number of items from each TEKS expectation that should be included on a test form. Verifying that test forms include the correct number of items from each TEKS expectation is a straightforward matter of counting items and matching blueprint percentages. These processes are summarized in the Chapter 2 and Chapter 4 of the Technical Digest. Additionally, under Task 1 of this report, we reviewed the 2016 STAAR forms and verified that the item content on each form matches those specified in the blueprint.

3.2. Build reliability expectations into test forms.

The IRT Rasch Model used by TEA to convert points for individual items into reported test scores drives the statistical considerations for test form construction. Basically, each assessment should have an array of items with varying degrees of difficulty, particularly around the score points that define differences between performance categories. This statistical consideration supports test reliability, particularly as computed by the concept of CSEM. TEA provided HumRRO with documentation on the statistical criteria used for test construction. These criteria specified the following: (a) include items with wide range of item difficulties, (b) exclude items that are too hard or too easy, and (c) avoid items with low item total correlations, which would indicate an item does not relate highly to other items on the test. Appendix B of the Technical Digest[17] shows acceptable CSEM for the 2015 test scores, and the projected CSEM estimates reported in Task 2 provide evidence that the test building process has adequately built reliability expectations into the test forms.

4. Administer Tests

In order for students' scores to have the same meaning, test administration must be consistent across students when scores are being interpreted within a given year and they must be consistent across years when scores are being interpreted as achievement gains across years. TEA provides instructions to all personnel involved in administering tests to students through test administration manuals.[18] The documentation provided by TEA is extensive, and sufficient time must be allocated for administrator preparation. To the extent that test administrators adequately prepare for the test administration and consistently follow the instructions provided by TEA, there is assurance that scores have the same meaning within a given year and across years.

5. Create Test Scores

Tests are administered each spring to students with the intent of measuring what a student knows and can do in relation to a specific grade and subject. The processes described above result in the creation of test forms. Students' responses to items on a given test are accumulated to produce a test score that is used to provide feedback on what a student knows and can do. The following procedures are used to create test scores.

[17] http://tea.texas.gov/Student_Testing_and_Accountability/Testing/Student_Assessment_Overview/Technical_Digest_2014-2015/

[18] http://tea.texas.gov/student.assessment/staar/manuals/

5.1. Conduct statistical item reviews.

Statistical item reviews are conducted for both field test items and then again for operational items. Chapter 3 of the Technical Digest lists standard items analyses, including p-values, item-total correlations, Rasch data and item graphs, and differential item functioning (DIF) analyses. These are typical statistics used for reviewing items and ensuring the items are functioning as expected.

5.2. Equate to synchronize scores across years.

Items used to compute grade/subject test scores are changed from one year to the next so that instruction does not become concentrated on particular test items. While tests across years are targeting the same blueprints and therefore should have equivalent content validity, tests across years may not be exactly equivalent in terms of the difficulty of the items. This creates a numerical issue for maintaining consistency in score meaning across years. This issue is solved using procedures that are typically referred to as equating. The solution involves placing items on the test form that have an established history. The difficulties of those equating items can be used to assess the difficulties of new items using well-established IRT processing, as described in the Technical Digest. Applying the results yields test scores that become numerically equivalent to prior years' scores. The one hurdle that, at times, must be addressed in this equating process is drift in an item. Drift is a detectable change in the difficulty of an item (for example, increased media attention of a specific topic area may make an item easier compared to the prior year). STAAR equating specifications detail one method for reviewing item drift. HumRRO is familiar with this method and believes that it will produce acceptable equating results.

5.3. Produce test form reliability statistics.

Chapter 4 of the Technical Digest adequately describes procedures for computing reliability, standard error of measurement, and conditional standard error of measurement. After the test is administered, this process is merely a post-hoc check on the extent to which adequate reliability was built into the test during form construction.

5.4. Produce final test scores.

Using the Rasch method for IRT, as implemented by Winsteps® (noted in the equating specifications document), involves reading Winsteps® tabled output to transform item total points to student ability estimates (i.e., IRT theta values). Theta values are on a scale that contains negative values so it is common practice to algebraically transform those values to a reporting scale. This is a simple linear transformation that does not impact validity or reliability.

Task 3 Conclusion

HumRRO reviewed the processes used to create STAAR test forms and the planned procedures for creating on-grade STAAR student scores. These scores are intended to be used to compare knowledge and skill achievements of students within and across years for a given grade/subject. TEA's test development process is consistent with best practices (Crocker & Algina, 1986) and includes a number of procedures that allow for the development of tests that measure and align with testable content.

HumRRO believes that these processes are adequate for developing tests that will yield scores that can be interpreted as representing what a student knows and can do. Further, the test development process ensures that each grade/subject test bears a strong association with on-grade curriculum requirements.

Overall Conclusion

In conclusion, HumRRO's independent evaluation finds support for the validity and reliability of the 2016 STAAR scores. Specifically:

Under Task 1, we identified evidence of the content validity of the assessments. The content review consisted of rating the alignment of each item to the Texas Essential Knowledge and Skills (TEKS) student expectation the item was intended to measure. Overall, the content of the 2016 forms aligned with blueprints, and HumRRO reviewers determined that the vast majority of items were aligned with the TEKS expectations for grades 3 through 8 mathematics and reading, grades 5 and 8 science, grade 8 social studies, and grades 4 and 7 writing.

Our work associated with Task 2 provided empirical evidence of the projected reliability and standard error of measurement for the 2016 forms. The projected reliability and conditional standard error of measurement (CSEM) estimates were all acceptable. Assuming the 2016 students' scores will have a similar distribution as the 2015 scores and assuming similar item functioning, the reliability and CSEM estimates based on 2016 student data should be similarly acceptable.

Finally, under Task 3, we reviewed the documentation of the test construction and scoring processes. Based on HumRRO's 20 years of experience in student achievement testing and 30 years of experience in high-stakes test construction, the processes used to construct the 2016 tests and the proposed methods for scoring the 2016 test are consistent with industry standards and support the development of tests that measure the knowledge and skills outlined in the content standards and test blueprint. The processes allow for the development of tests that yield valid and reliable assessment scores.

References

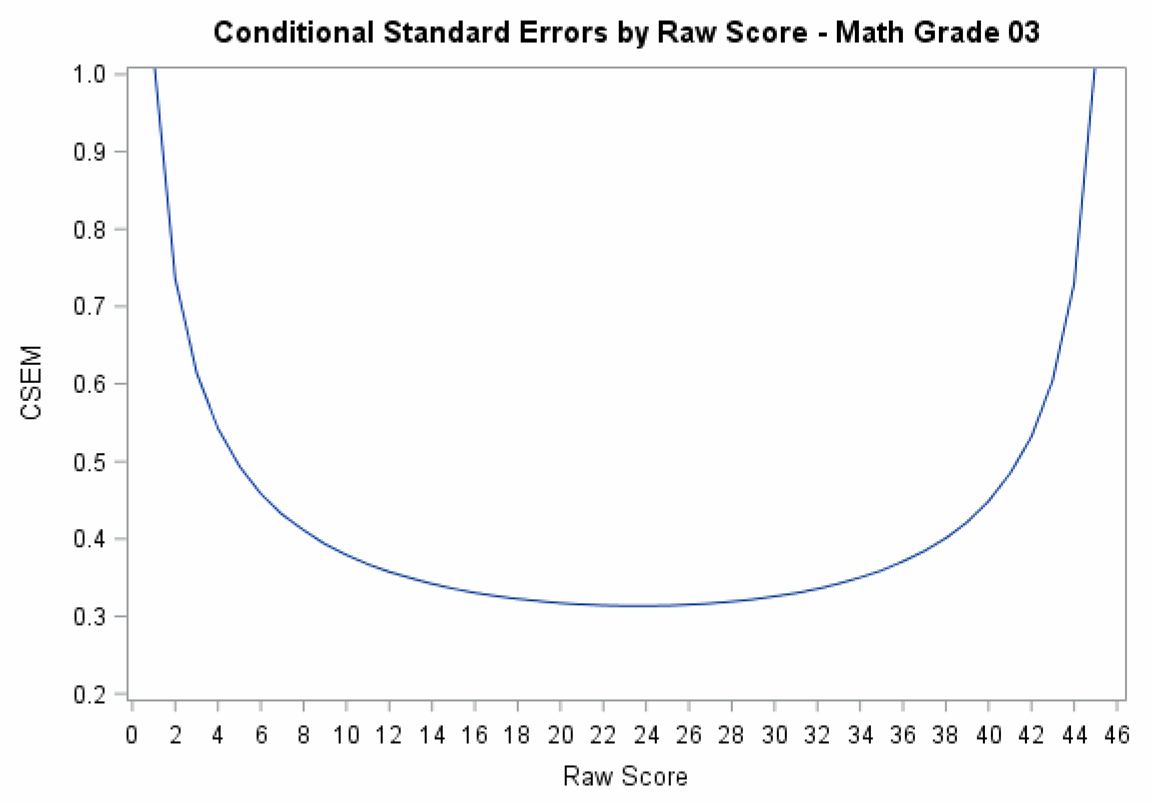
Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: CBS College Publishing.

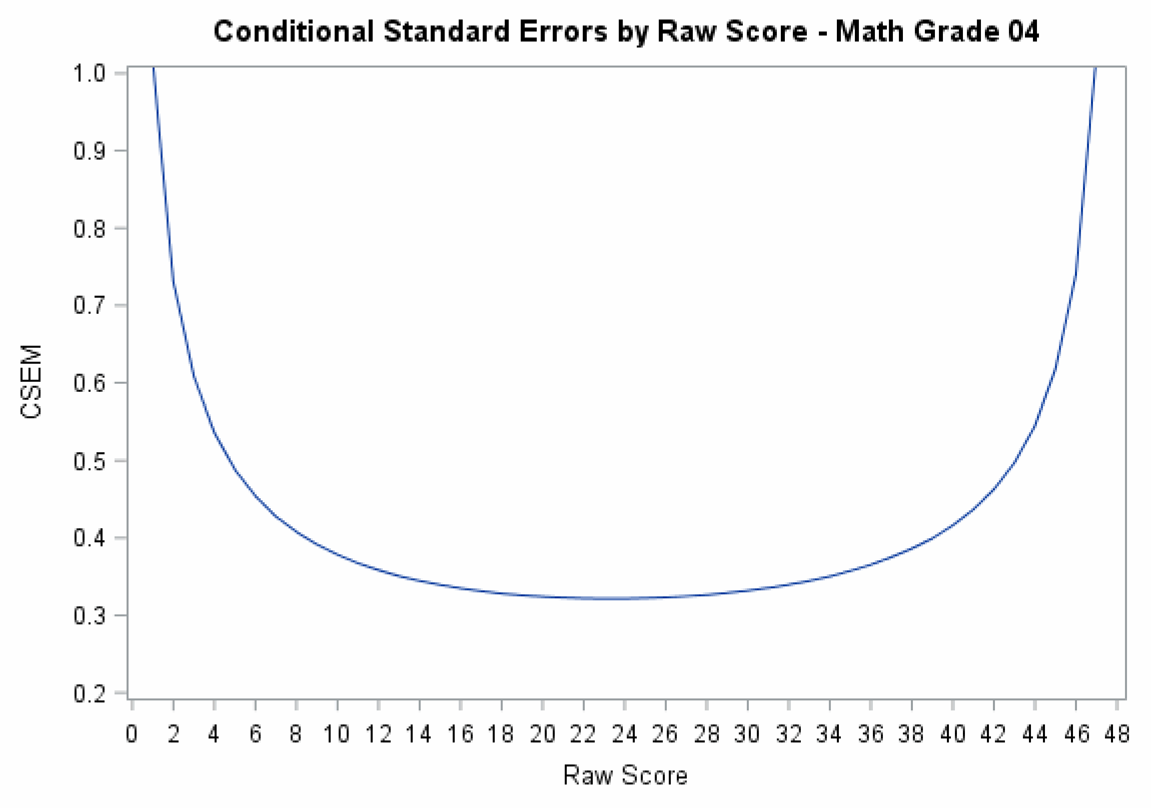
Kolen, M. J., Zang, L., & Hanson, B. A. (1996). Conditional standard errors of measurement for scale scores Using IRT. *Journal of Educational Measurement, 33*(2), 129-140.

Linacre, J. M. (2016). Winsteps® Rasch measurement computer program. Beaverton, Oregon: Winsteps.com

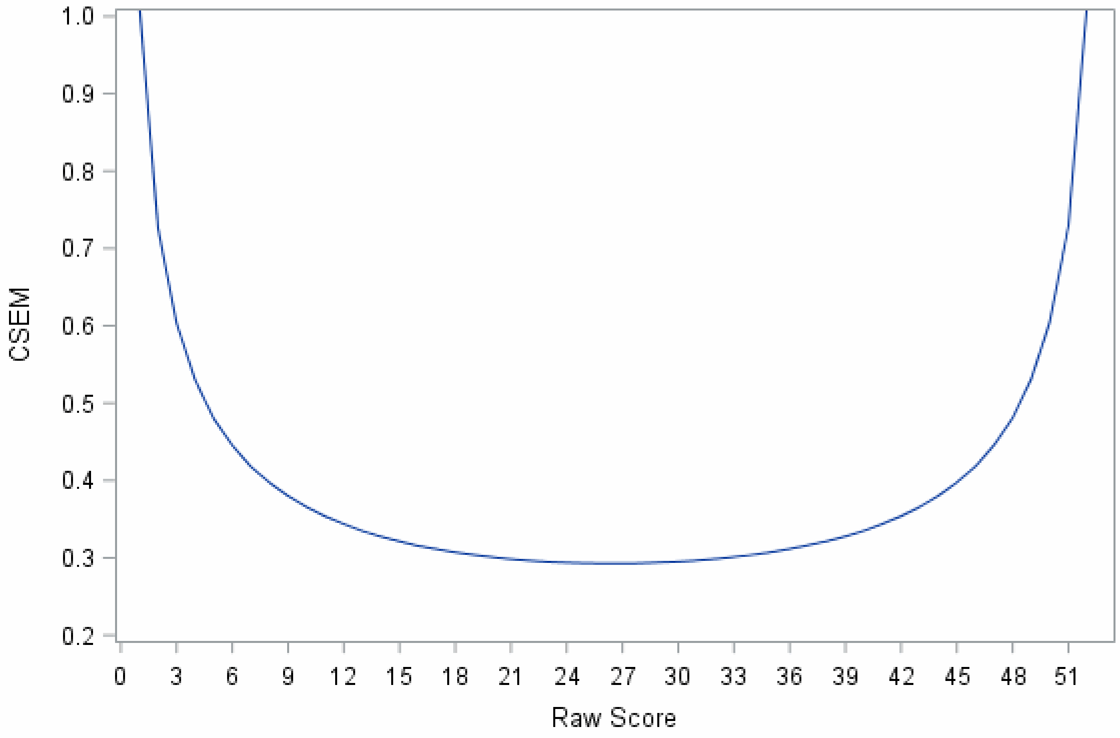
Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.

Appendix A: Conditional Standard Error of Measurement Plots

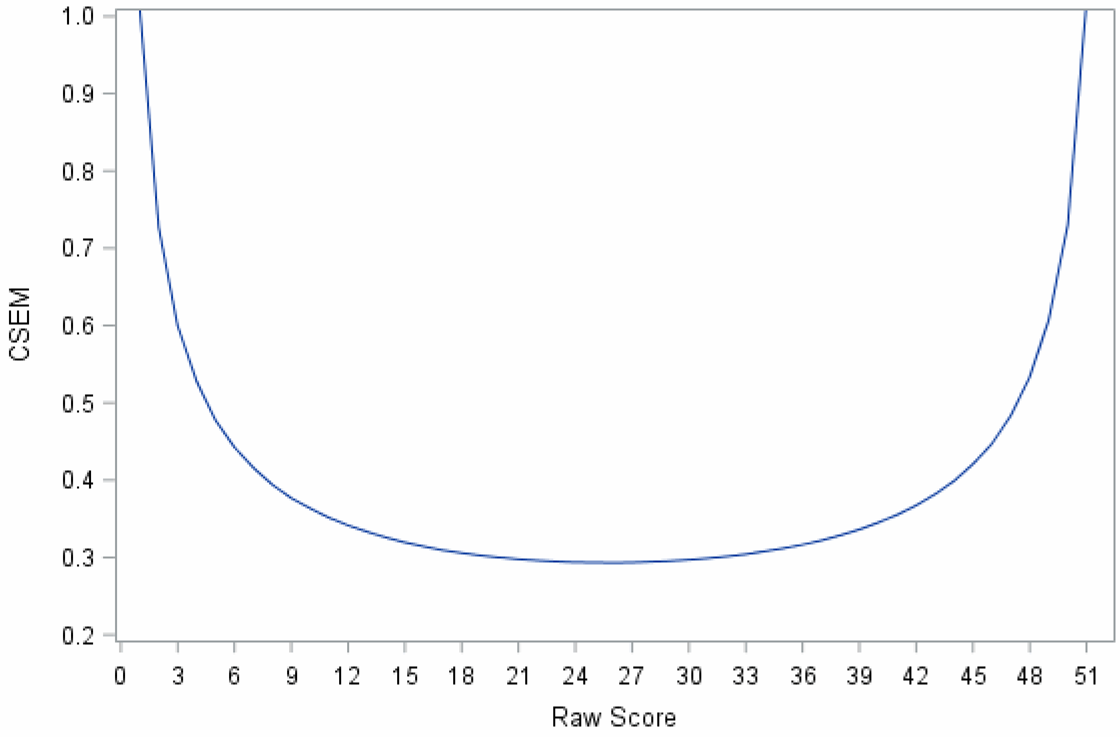


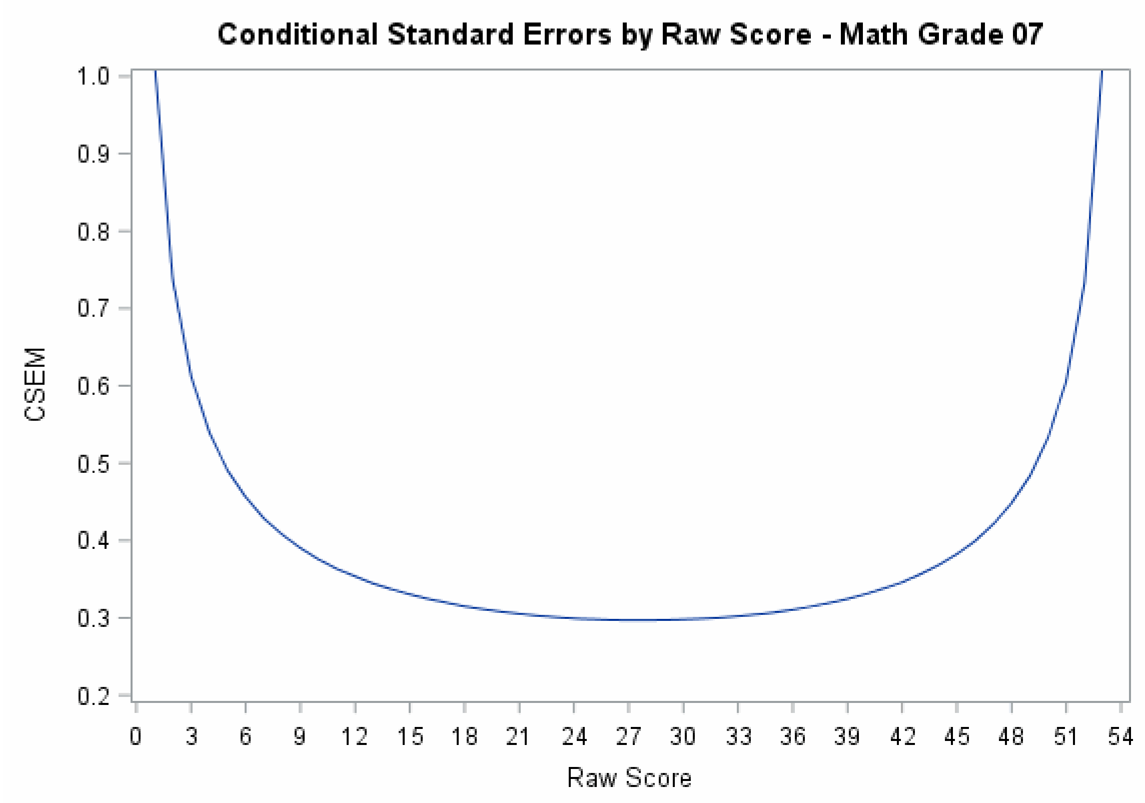


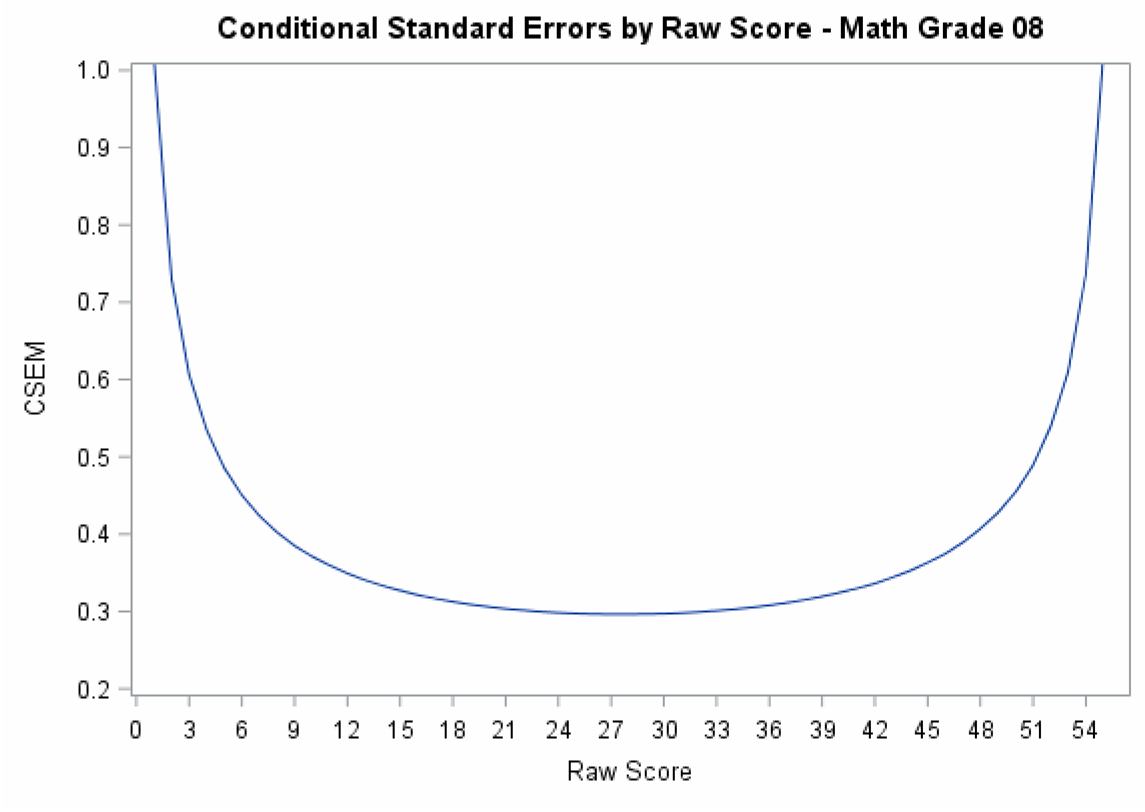
Conditional Standard Errors by Raw Score - Math Grade 05

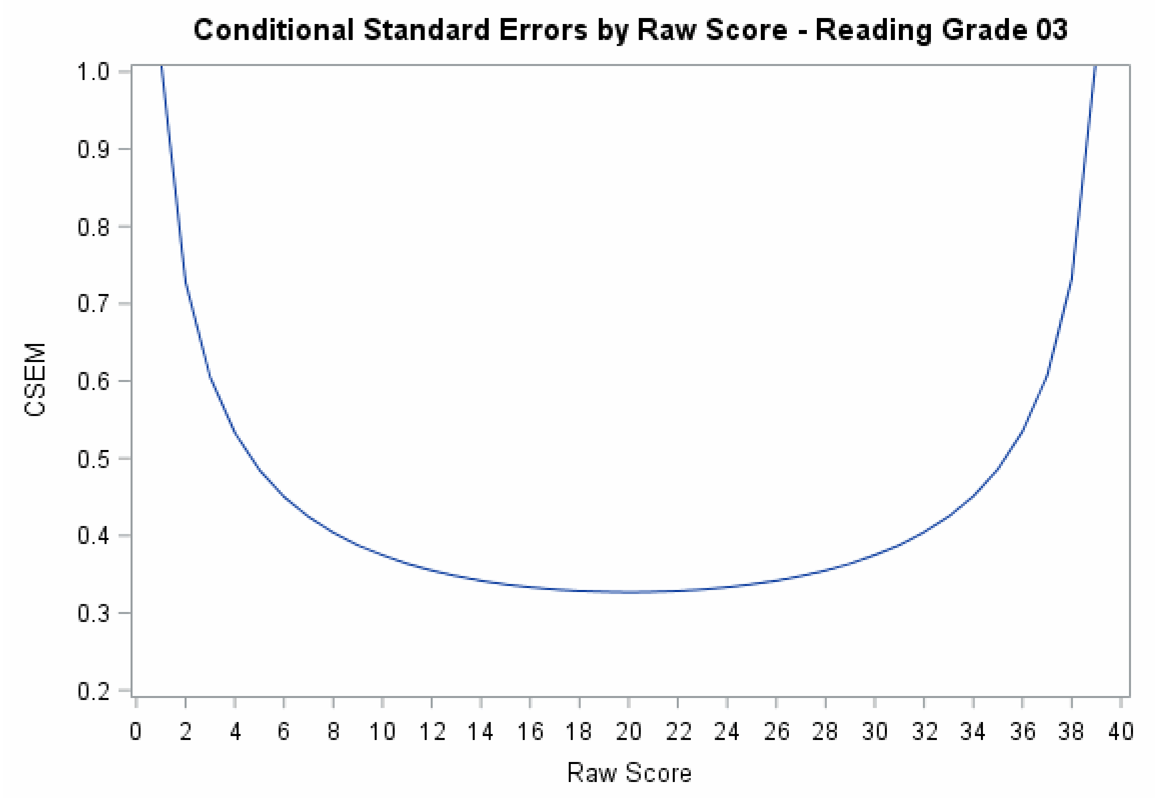


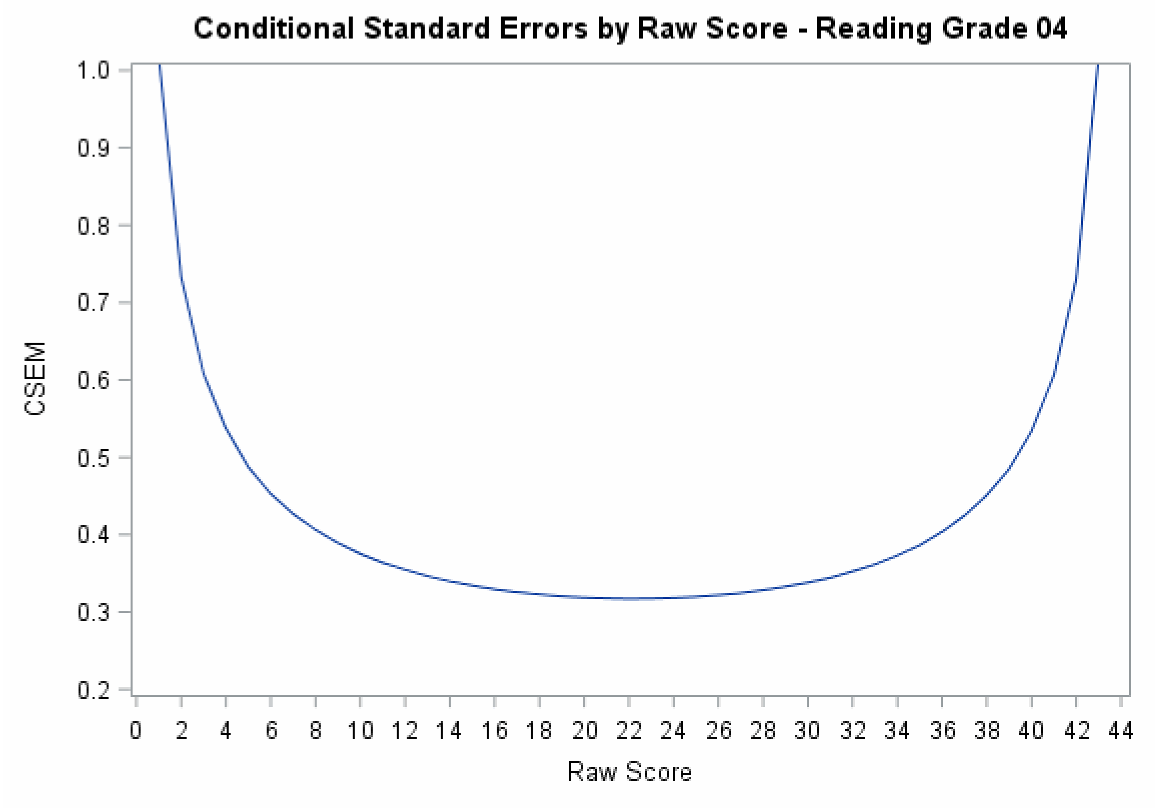
Conditional Standard Errors by Raw Score - Math Grade 06

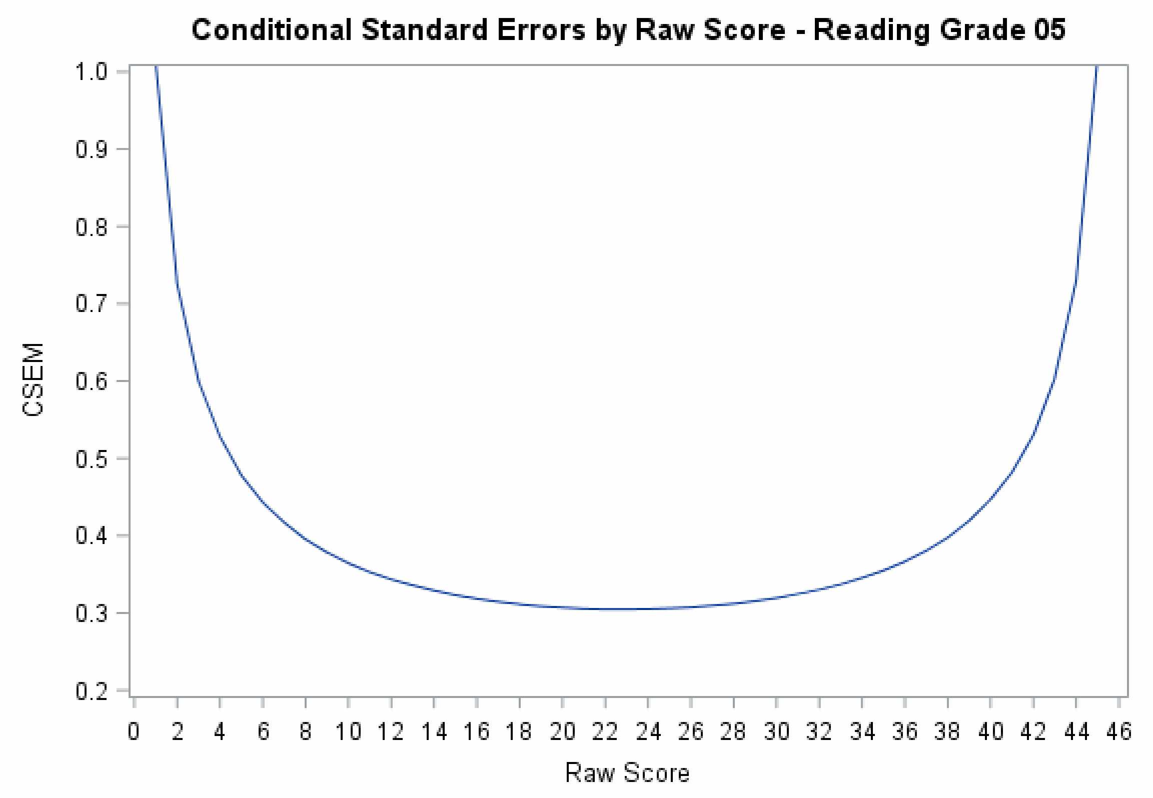


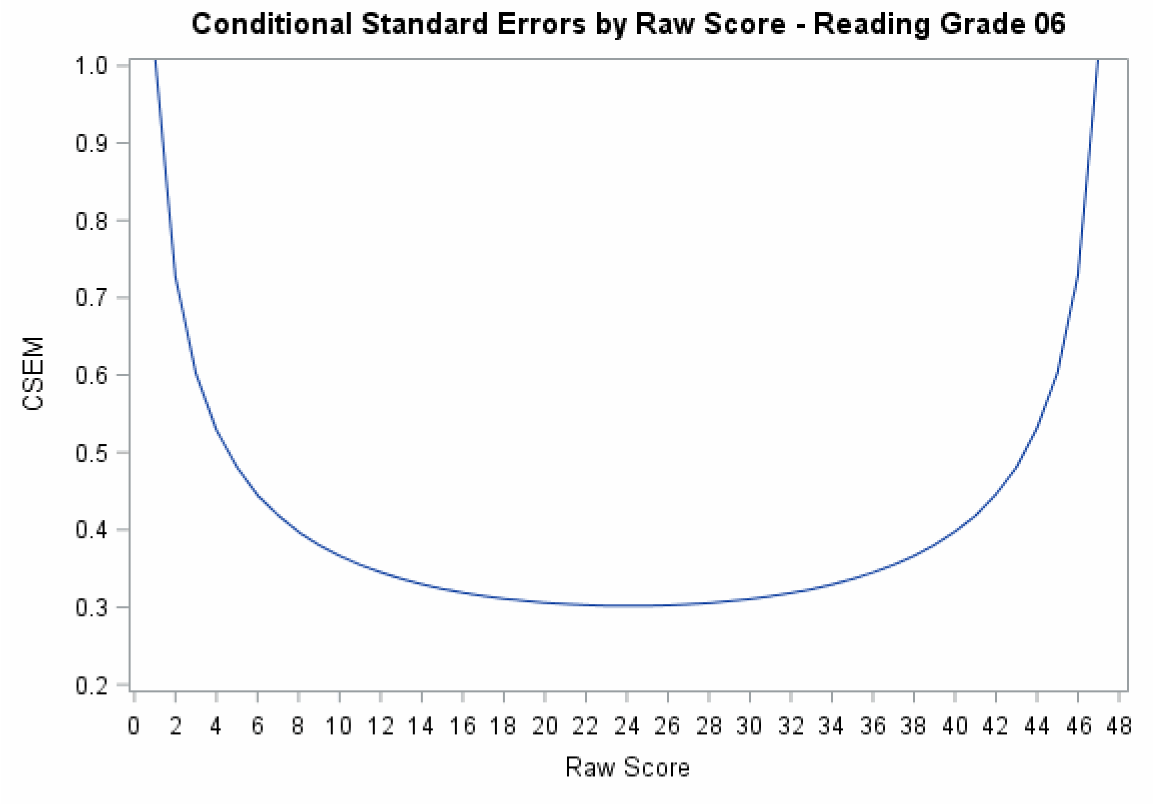

Conditional Standard Errors by Raw Score - Math Grade 07
CSEM
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0
3
6
9
12
15
18
21
24
27
30
33
36
39
42
45
48
51
54
Raw Score

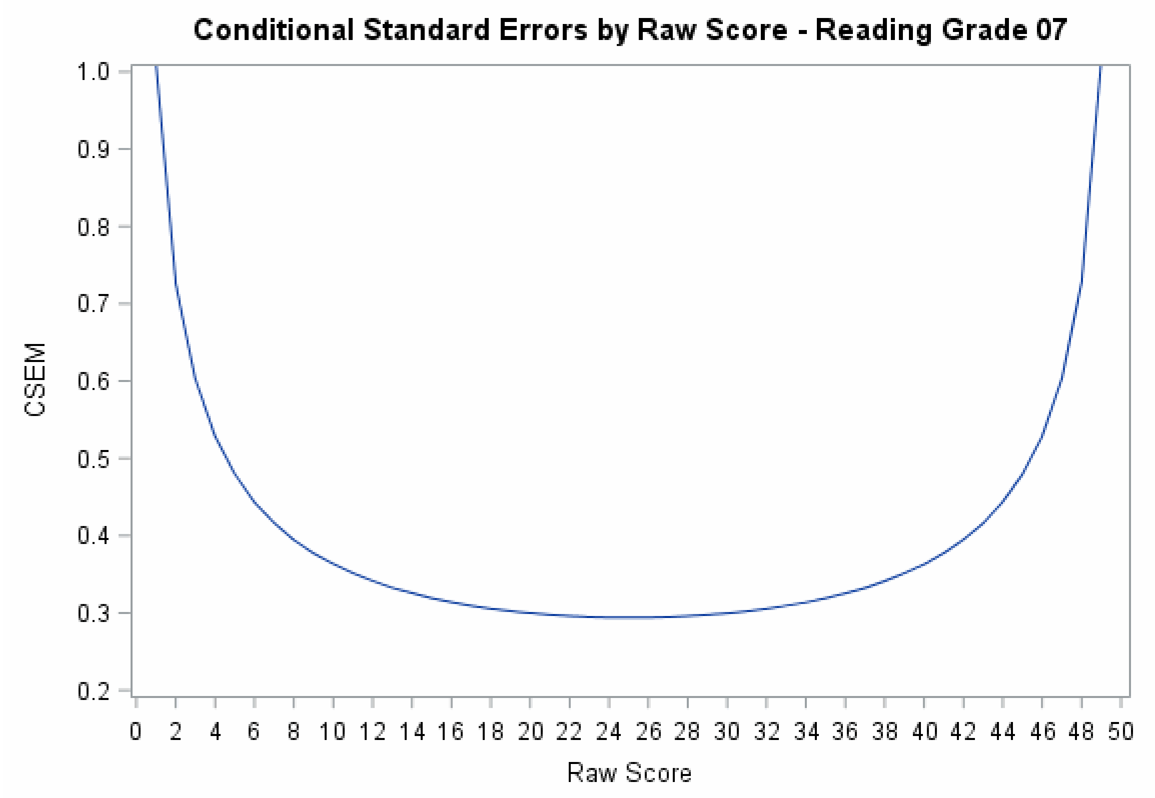

Conditional Standard Errors by Raw Score - Math Grade 08
CSEM
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0
3
6
9
12
15
18
21
24
27
30
33
36
39
42
45
48
51
54
Raw Score


Conditional Standard Errors by Raw Score - Reading Grade 03
CSEM
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0
2
4
6
8
10
12
14
16
18
20
22
24
26
28
30
32
34
36
38
40
Raw Score

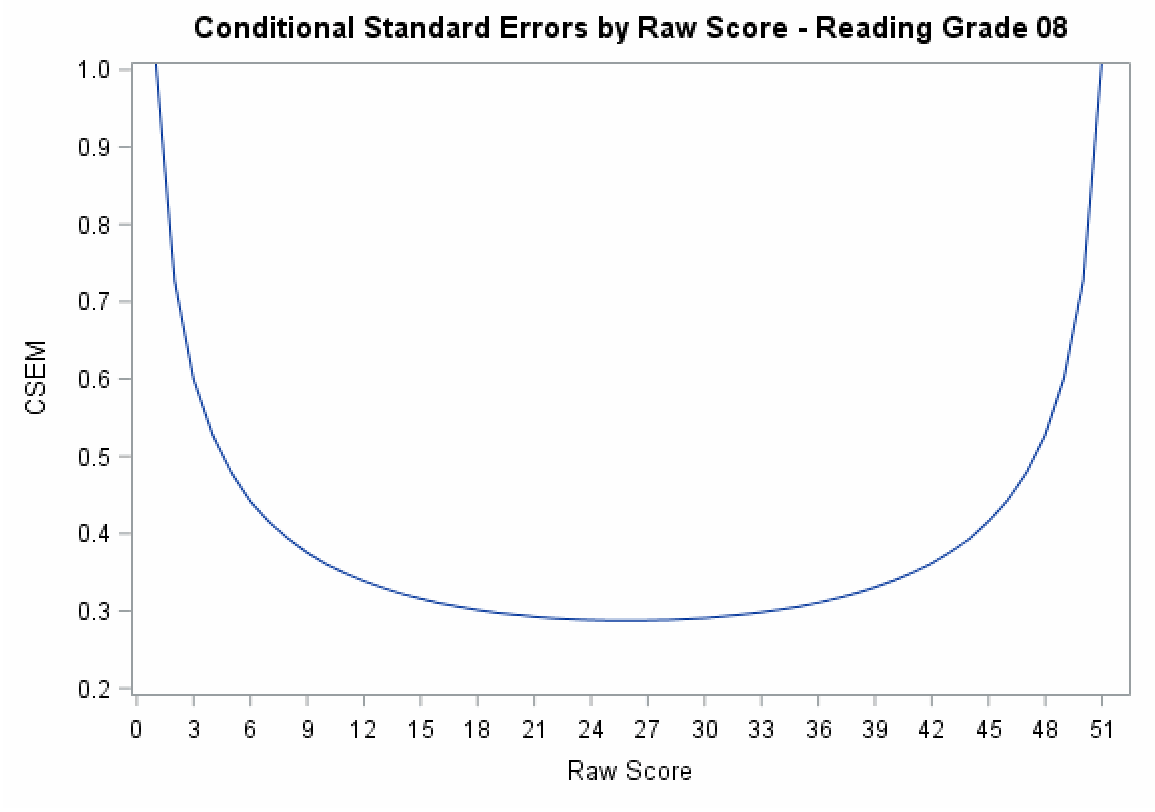

Conditional Standard Errors by Raw Score - Reading Grade 04
CSEM
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0
2
4
6
8
10
12
14
16
18
20
22
24
26
28
30
32
34
36
38
40
42
44
Raw Score


Conditional Standard Errors by Raw Score - Reading Grade 05
CSEM
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0 2 4 6 8 10 12 14 16 18 20 22 24 26 28 30 32 34 36 38 40 42 44 46
Raw Score


Conditional Standard Errors by Raw Score - Reading Grade 06
CSEM
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0 2 4 6 8 10 12 14 16 18 20 22 24 26 28 30 32 34 36 38 40 42 44 46 48
Raw Score

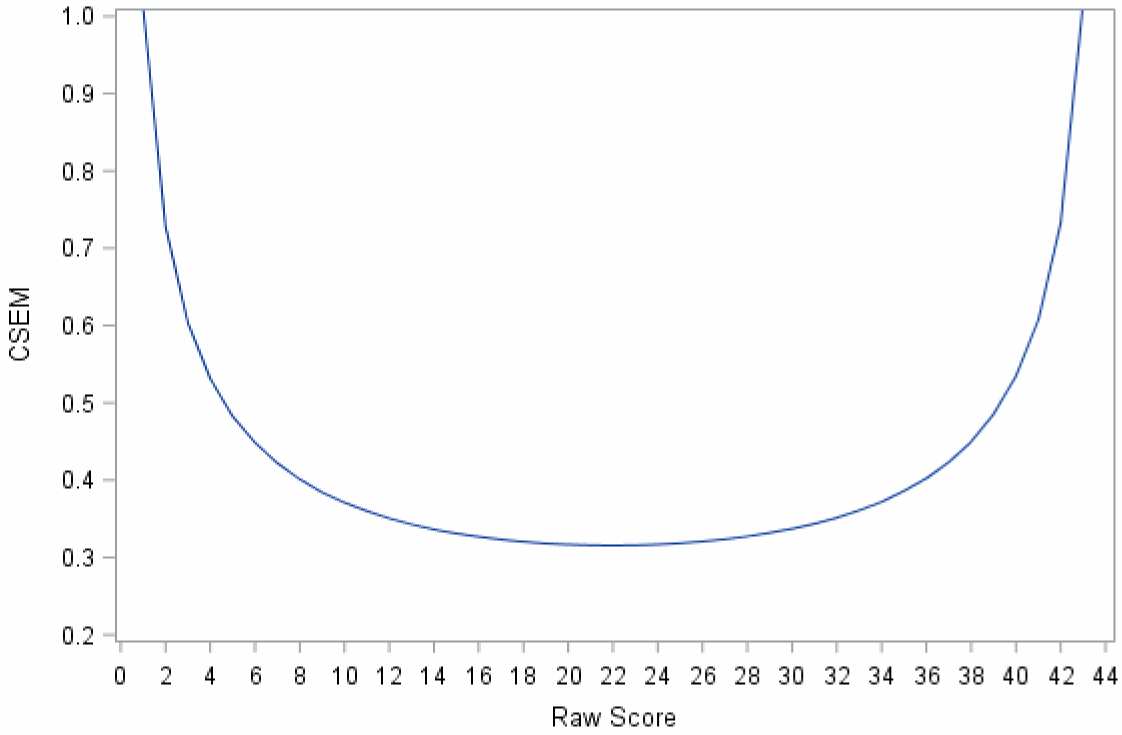


Conditional Standard Errors by Raw Score - Reading Grade 07

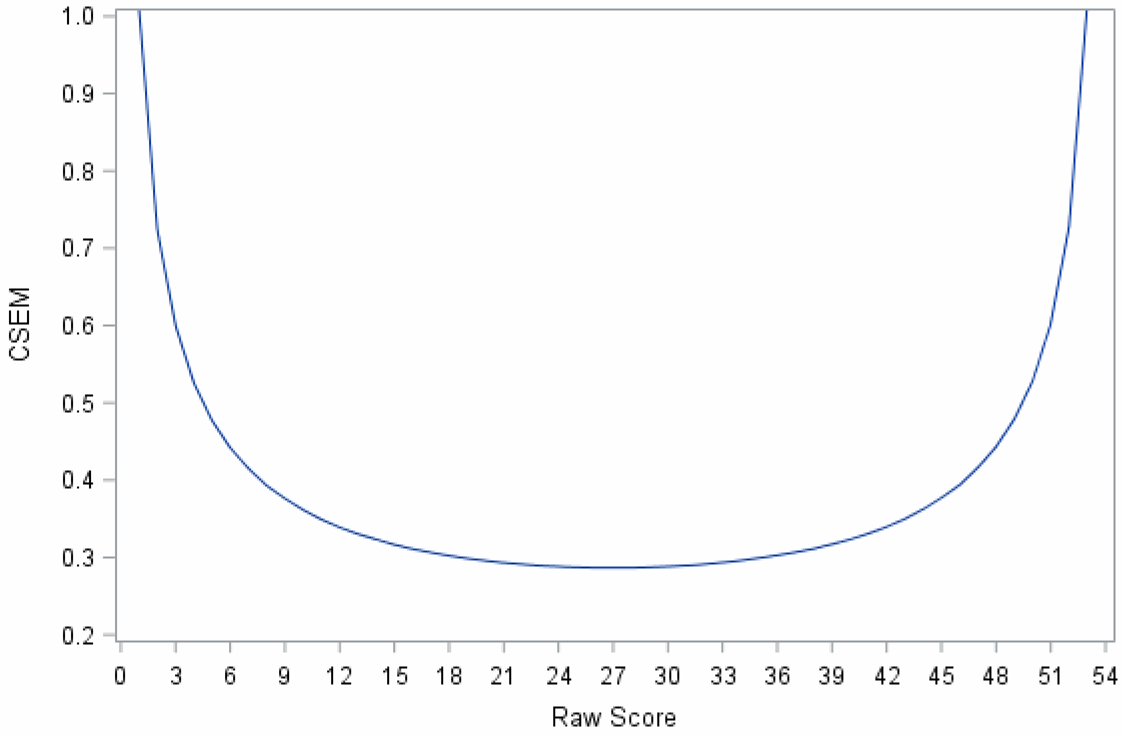


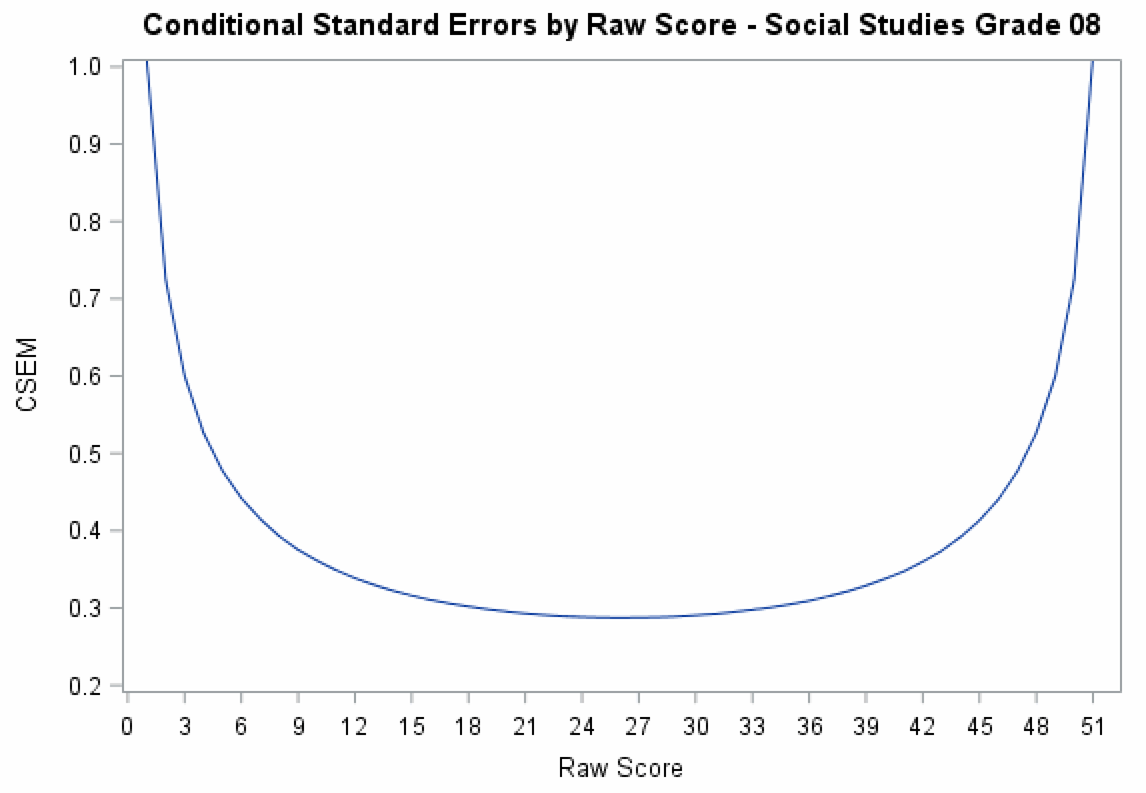
Conditional Standard Errors by Raw Score - Reading Grade 08

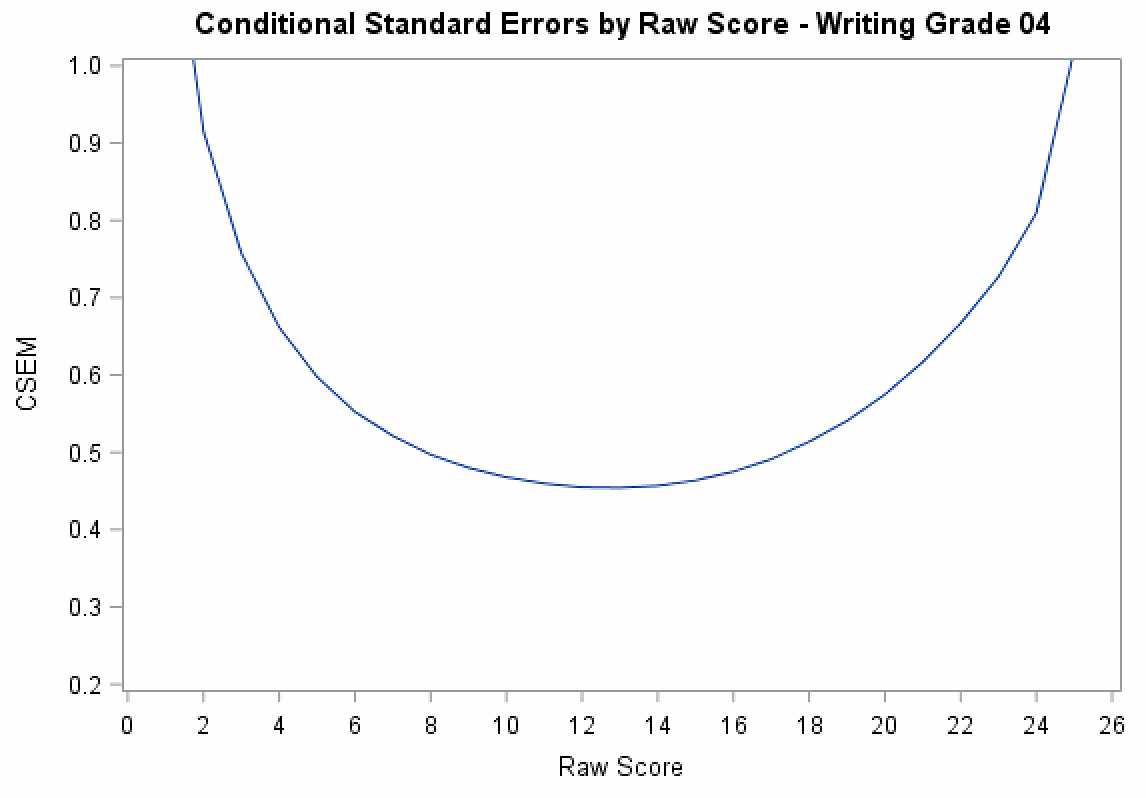
Conditional Standard Errors by Raw Score - Science Grade 05

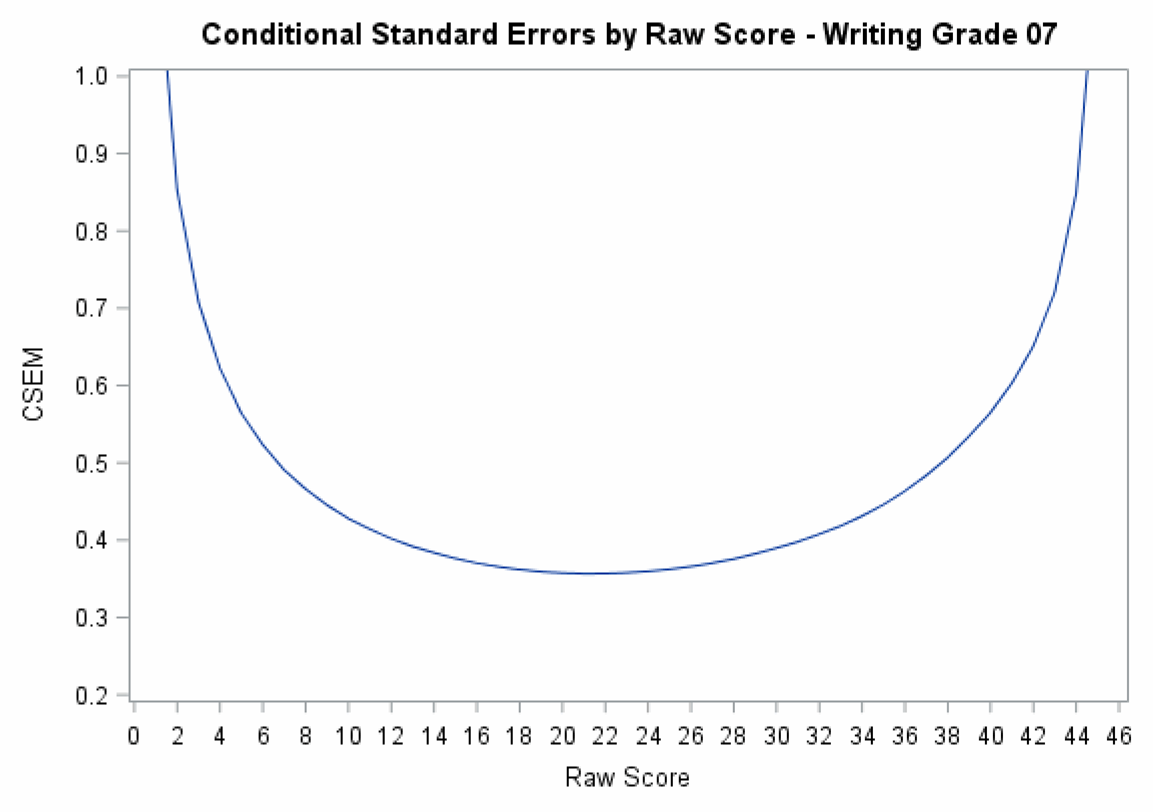


Conditional Standard Errors by Raw Score - Science Grade 08




Conditional Standard Errors by Raw Score - Social Studies Grade 08


Conditional Standard Errors by Raw Score - Writing Grade 04


Conditional Standard Errors by Raw Score - Writing Grade 07
CSEM
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0
2
4
6
8
10
12
14
16
18
20
22
24
26
28
30
32
34
36
38
40
42
44
46
Raw Score